Vol. 277, No. 2, May 28, 1993
Completing Volume 277

ISSN 0003-2670

Special Issue
PAPERS PRESENTED AT THE FIFTH INTERNATIONAL CONFERENCE ON CHEMOMETRICS IN ANALYTICAL CHEMISTRY, MONTREAL, CANADA, JULY 14–17, 1992

ANALYTICA CHIMICA ACTA

An international journal devoted to all branches of analytical chemistry

ELSEVIER

ANALYTICA CHIMICA ACTA

Scope. *Analytica Chimica Acta* publishes original papers, preliminary communications and reviews dealing with every aspect of modern analytical chemistry. Reviews are normally written by invitation of the editors, who welcome suggestions for subjects. Preliminary communications of important urgent work can be printed within four months of submission, if the authors are prepared to forego proofs.

Submission of Papers

Americas

Prof. Harry L. Pardue
Department of Chemistry
1393 BRWN Bldg, Purdue University
West Lafayette, IN 47907–1393
USA
Tel:(+1-317) 494 5320
Fax:(+1-317) 496 1200

Computer Techniques

Prof. J.T. Clerc
Universität Bern
Pharmazeutisches Institut
Baltzerstrasse 5, CH-3012 Bern
Switzerland
Tel:(+41-31) 654171
Fax:(+41-31) 654198

Other Papers

Prof. Alan Townshend
Department of Chemistry
The University
Hull HU6 7RX
Great Britain

Tel:(+44-482) 465027
Fax:(+44-482) 466410

Prof. Willem E. van der Linden
Laboratory for Chemical Analysis
Department of Chemical Technology
Twente University of Technology
P.O. Box 217, 7500 AE Enschede
The Netherlands
Tel:(+31-53) 892629
Fax:(+31-53) 356024

Prof. Paul Worsfold
Dept. of Environmental Sciences
University of Plymouth
Plymouth PL4 8AA
Great Britain

Tel:(+44-752) 233006
Fax:(+44-752) 233009

Submission of an article is understood to imply that the article is original and unpublished and is not being considered for publication elsewhere. *Anal. Chim. Acta* accepts papers in English only. There are no page charges. Manuscripts should conform in layout and style to the papers published in this issue. See inside back cover for "Information for Authors".

Publication. *Analytica Chimica Acta* appears in 14 volumes in 1993. The subscription price for 1993 (Vols. 267–280) is Dfl. 4214.00 plus Dfl. 462.00 (p.p.h.) (total approx. US$ 2672.00). *Vibrational Spectroscopy* appears in 2 volumes in 1993. The subscription price for *Vibrational Spectroscopy* (Vols. 4 and 5) is Dfl. 700.00 plus Dfl. 66.00 (p.p.h.) (total approx. US$ 437.75). The price of a combined subscription (*Anal. Chim. Acta* and *Vib. Spectrosc.*) is Dfl. 4592.00 plus Dfl. 528.00 (p.p.h.) (total approx. US$ 2925.75). All earlier volumes (Vols. 1–266) except Vols. 23 and 28 are available at Dfl. 259.50 (US$ 148.25), plus Dfl. 18.00 (US$ 10.25) p.p.h., per volume. The Dutch guilder price is definitive. The U.S. dollar price is subject to exchange-rate fluctuations and is given only as a guide. Subscriptions are accepted on a prepaid basis only, unless different terms have been previously agreed upon.
Our p.p.h. (postage, packing and handling) charge includes surface delivery of all issues, except to subscribers in the U.S.A., Canada, Australia, New Zealand, China, India, Israel, South Africa, Malaysia, Thailand, Singapore, South Korea, Taiwan, Pakistan, Hong Kong, Brazil, Argentina and Mexico, who receive all issues by air delivery (S.A.L.–Surface Air Lifted) at no extra cost. For Japan, air delivery requires 25% additional charge of the normal postage and handling charge; for all other countries airmail and S.A.L. charges are available upon request.

Subscription orders. Subscription orders can be entered only by calendar year and should be sent to: Elsevier Science Publishers B.V., Journals Department, P.O. Box 211, 1000 AE Amsterdam, The Netherlands. Tel: (+31-20) 5803 642, Telex: 18582, Telefax: (+31-20) 5803598, to which requests for sample copies can also be sent. Claims for issues not received should be made within six months of publication of the issues. If not they cannot be honoured free of charge. Readers in the U.S.A. and Canada can contact the following address: Elsevier Science Publishing Co. Inc., Journal Information Center, 655 Avenue of the Americas, New York, NY 10010, U.S.A. Tel: (+1-212) 6333750, Telefax: (+1-212) 6333990, for further information, or a free sample copy of this or any other Elsevier Science Publishers journal.
Advertisements. Advertisement rates are available from the publisher on request.

Detailed "Instructions to Authors" for *Analytica Chimica Acta* was published in Volume 256, No. 2, pp. 373–376. Free reprints of the "Instructions to Authors" of *Analytica Chimica Acta* and *Vibrational Spectroscopy* are available from the Editors or from: Elsevier Science Publishers B.V., P.O. Box 330, 1000 AH Amsterdam, The Netherlands. Telefax: (+31-20) 5862845.

US mailing notice – Analytica Chimica Acta (ISSN 0003-2670) is published biweekly by Elsevier Science Publishers (Molenwerf 1, Postbus 211, 1000 AE Amsterdam). Annual subscription price in the USA US$ 2672.00 (subject to change), including air speed delivery. Second class postage paid at Jamaica, NY 11431. *USA Postmasters*: Send address changes to Anal. Chim. Acta, Publications Expediting, Inc., 200 Meacham Av., Elmont, NY 11003. Airfreight and mailing in the USA by Publication Expediting.

ANALYTICA CHIMICA ACTA

An international journal devoted to all branches of analytical chemistry

(Full texts are incorporated in CJELSEVIER, a file in the Chemical Journals Online database available on STN International; Abstracted, indexed in: Aluminum Abstracts; Anal. Abstr.; Biol. Abstr.; BIOSIS; Chem. Abstr.; Curr. Contents Phys. Chem. Earth Sci.; Engineered Materials Abstracts; Excerpta Medica; Index Med.; Life Sci.; Mass Spectrom. Bull.; Material Business Alerts; Metals Abstracts; Sci. Citation Index)

VOL. 277 NO. 2 CONTENTS MAY 28, 1993

Papers presented at the Fifth International Conference on Chemometrics in Analytical Chemistry, Montreal, Canada, July 14–17, 1992

Foreword 163

General Applications of Chemometrics

The parsimony principle applied to multivariate calibration
M.B. Seasholtz and B. Kowalski (Seattle, WA, USA) 165

Colored information from a black box? Validation and evaluation of neural networks
G. Kateman and J.R.M. Smits (Nijmegen, Netherlands) 179

FLIN: Fuzzy Linear Interpolating Network
P. De B. Harrington and B.W. Pack (Athens, OH, USA) 189

Some novel methods based on recursive optimal estimation. Applications to analytical chemistry
B. Yu, M. Li, A. Liu, Z. Li (Changsha, China), L. Shi and Z. Pan (Beijing, China) 199

Multivariate decision and detection limits
A. Singh (Las Vegas, NV, USA) 205

Some graphical aids for univariate exploratory data analysis
J. Militký (Liberec, Czech Republic) and M. Meloun (Pardubice, Czech Republic) 215

Modelling

Developing models for infinite dilution activity coefficients using factor analysis methods
R.B. Poe, S.C. Rutan (Richmond, VA, USA), M.J. Hait, C.A. Eckert (Atlanta, GA, USA) and P.W. Carr (Minneapolis, MN, USA) 223

DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures
S. Wold, J. Jonsson, M. Sjöström, M. Sandberg and S. Rännar (Umeå, Sweden) 239

Superfund site characterization using non-parametric variogram modeling
A.K. Singh, M.M.A. Ananda and A.R. Sparks (Las Vegas, NV, USA) 255

Use of the mean quadratic error of prediction for the construction of biased linear models
J. Militký (Liberec, Czech Republic) and M. Meloun (Pardubice, Czech Republic) 267

Neural Networks

Data processing using neural networks
T.B. Blank and S.D. Brown (Newark, DE, USA) 273

Artificial neural networks as a multivariate calibration tool: modeling the Fe–Cr–Ni system in x-ray fluorescence spectroscopy
A. Bos, M. Bos and W.E. Van der Linden (Enschede, Netherlands) 289

Applications in Analytical Spectroscopy

Hypermedia tools for the interpretation of mass spectra
T. Brodmeier, A. Gloor, M. Cadisch, R. Bürgin and E. Pretsch (Zurich, Switzerland) 297

Pattern recognition studies of tandem mass spectra
D. Swain, W.J. Dunn, III (Chicago, IL, USA) and R.E. Talaat (Columbia, MO, USA) 305

(Continued overleaf)

Contents (continued)

HIPS, a hybrid self-adapting expert system for nuclear magnetic resonance spectrum interpretation using genetic algorithms
R. Wehrens, C. Lucasius, L. Buydens and G. Kateman (Nijmegen, Netherlands) 313

Applications in Chromatography
Information theory of chromatography and titration
Y. Hayashi and R. Matsuda (Tokyo, Japan) 325
Taxonomy of *Amanita* mushrooms by pattern recognition of amino acid chromatographic data
R.L. White, P.D. Wentzell, M.A. Beasy (Halifax, Canada), D.S. Clark and D.W. Grund (Wolfville, Canada) 333

Applications in Environmental Analysis
Applying and developing receptor models to the 1990 El Paso air data: a look at receptor modeling with uncharacterized sources and graphical diagnostics
C.H. Spiegelman and S. Dattner (College station, TX, USA) 347
Gas chromatography–pattern recognition techniques in pollution monitoring
B.K. Lavine, A. Stine (Potsdam, NY, USA) and H.T. Mayfield (Tyndall AFB, FL, USA) 357
Potential source contribution function analysis and source apportionment of sulfur species measured at Rubidoux, CA during the Southern California Air Quality Study, 1987
N. Gao, M.-D. Cheng and P.K. Hopke (Potsdam, NY, USA) 369
Environmental applications of combined multidimensional gas chromatography–infrared–mass spectrometry
K.A. Krock and C.L. Wilkins (Riverside, CA, USA) 381
Atmospheric chemometrics for identification of trace element sources in precipitation
R.J. Vong (Corvallis, OR, USA) 389
Empirical Bayes estimation in factor analysis for aerosol mass apportionment
L.J. Gleser and H. Yang (Pittsburgh, PA, USA) 405
Receptor models to study groundwater contamination
E.J. Baum (Allendale, MI, USA) 421
Optimal frequency of quality control analyses in environmental measurement
M.J. Miah, F.C. Garner, M.A. Stapanian and G.A. Laing (Las Vegas, NV, USA) 431

Expert Systems
Development of an expert system for selection of experimental designs
R.A. Olivero (Las Vegas, NV, USA), S. Seshadri (Sugarland, TX, USA) and S.N. Deming (Houston, TX, USA) 441

Applications to Mixture Analysis
Designs for mixture and process variables applied in tablet formulations
C.A.A. Duineveld, A.K. Smilde and D.A. Doornbos (Groningen, Netherlands) 455
Prediction of mixture composition by chromatographic characterization, multivariate classification and partial least-squares regression, a comparison of methods
P. Kaufmann (Stockholm, Sweden) 467

Outliers and Intervals
Bayesian slippage test for detection of outlying sub-samples
A.K. Singh and A. Singh (Las Vegas, NV, USA) 473
Simple regression and outlier detection using the median method
M.O. Moen, K.J. Griffin and A.H. Kalantar (Edmonton, Canada) 477
Outlier detection by robust alternating regression
Å. Ukkelberg and O.S. Borgen (Trondheim, Norway) 489
An approach to interval estimation in partial least squares regression
A. Phatak, P.M. Reilly and A. Penlidis (Waterloo, Canada) 495
Bayesian confidence intervals for the product of three normal means
M.M.A. Ananda, A.K. Singh and G.T. Flatman (Las Vegas, NV, USA) 503

Author Index 511

SPECIAL ISSUE

PAPERS PRESENTED AT THE FIFTH INTERNATIONAL CONFERENCE ON CHEMOMETRICS IN ANALYTICAL CHEMISTRY, MONTREAL, CANADA, JULY 14–17, 1992

Analytica Chimica Acta, 277 (1993) 163
Elsevier Science Publishers B.V., Amsterdam

FOREWORD

It is with great pleasure to be able to write this editorial introducing this special issue of *Analytica Chimica Acta* in which the papers presented at the *Fifth International Conference on Chemometrics in Analytical Chemistry* (CAC'92) are published. This conference held between July 14–17, 1992 in Montreal was the first time one of the continuing chemometrics conferences had been moved from Europe to North America. There have been chemometrics symposia at the American Chemical Society and FACSS meetings, and the Environmental Protection Agency has held several conferences devoted specifically to environmental problems. However, this opportunity to hold the fifth in a series of successful CAC meetings was important for the continuing development and integration of chemometrics into North American chemistry. Although we can now see progress in the recognition by the chemical community of the utility of chemometrics in extracting information from chemical data, it is still not thought of as an integral part of chemical training nor as a fully accepted research specialty deserving of support by those agencies who regularly fund basic chemical scientific studies.

Over 90 people attended the meeting including a large number of our European colleagues. The scientific content of the meeting can best be judged by examination of the papers that follow. Virtually all of the presentations have been converted into manuscripts of very good quality. It was also a great pleasure to be able to participate in the awarding of the first *Elsevier Prize in Chemometrics* to Dr. Lutgarde Buydens of the University of Nijmegen.

Thus, I believe this meeting served its purpose in providing a useful forum for the exchange of ideas and is a starting point for major meetings on Chemometrics to be held in North America. It is hoped that the momentum gained from this meeting can be maintained. However, it will be necessary for others in the field to come forward and help lead the effort to develop a continuing series of such chemometrics meetings as have been common in Europe for almost a decade. I would hope that another similar gathering could be organized in 1995 or 1996.

Finally, I would be remiss if I did not again thank some of the many people who helped put the meeting together including Drs. Barry Lavine, Donald Scott and George Flatman, and Professor Steven Brown. These individuals did most of the solicitation of the papers we enjoyed hearing and now reading. Jan Lavine was a great asset in helping with many of the administrative tasks and Eleanor Hopke and Rose Casaleno were great at manning the registration desk. Financial support was provided by Clarkson University's School of Science and the U.S. Environmental Protection Agency through the Office of Exploratory Research, the Atmospheric Research and Exposure Assessment Laboratory, and the Environmental Monitoring Systems Laboratory. I hope that CAC'92 represents one of many more productive scientific conferences in chemometrics held in North America.

Philip K. Hopke

Analytica Chimica Acta, 277 (1993) 165–177
Elsevier Science Publishers B.V., Amsterdam

The parsimony principle applied to multivariate calibration

Mary Beth Seasholtz and Bruce Kowalski
Laboratory for Chemometrics, BG-10, University of Washington, Seattle, WA 98195 (USA)

(Received 2nd July 1992)

Abstract

The general principle of parsimonious data modeling states that if two models in some way adequately model a given set of data, the one that is described by a fewer number of parameters will have better predictive ability given new data. This concept is of interest in multivariate calibration since several new non-linear modeling techniques have become available. Three such methods are neural networks, projection pursuit regression (PPR) and multivariate adaptive regression splines (MARS). These methods, while capable of modeling non-linearities, typically have very many parameters that need to be estimated during the model building phase. The biased calibration methods, principal components regression (PCR) and partial least squares (PLS) are linear methods and so may not as efficiently describe some types of non-linearities, however have comparably very few parameters to be estimated. It is therefore of interest to study the parsimony principle formally in order to understand under what circumstances the various methods are appropriate. In this paper, the mathematical theory of parsimonious data modeling is presented. The assumptions made in the theory are shown to hold for multivariate calibration methods. This theory is used to provide a procedure for selecting the most parsimonious model structure for a given calibration application.

Keywords: Multivariate calibration; Partial least squares; Principal component analysis; Chemometrics; Model selection; Multivariate adaptive repression splines; Neural networks; Parsimony principle; Projection pursuit regression

There are now available a variety of multivariate calibration methods [1–5]. It is not always clear, however, which method is optimal for any particular application. Since the primary purpose for developing a model in analytical chemistry is for prediction of future sample properties, one approach for selecting an optimal method is to calculate a prediction statistic for each calibration technique and select the method with the smallest prediction error. A widely used validation statistic given a limited set of data is leave-one-out cross validation [6], which is a nearly unbiased estimate of the true prediction error [7]. However, the variance of this estimate is quite large [7], and so it may not be a suitable statistic for method comparison.

Multivariate calibration methods span a wide range of model complexity, and so can also be compared in terms of how parsimonious or simple they are. A general rule of thumb for parsimonious data modeling simply states that of two meaningful models, the one that is described by fewer parameters will have better predictive ability given new data. Intuitively this is appealing since if there are fewer parameters, there is perhaps less propagation of error from the data into the parameter estimates, and so overfitting will be minimized. This principle was seen to hold true in Refs. 8 and 9, where models built with selected wavelengths had smaller prediction er-

Correspondence to: B. Kowalski, Laboratory for Chemometrics, BG-10, University of Washington, Seattle, WA 98195 (USA).

rors than if all the available wavelengths were used. It is of interest therefore to study this 'parsimony principle' formally in order to understand under what circumstances the various calibration methods are appropriate.

In this paper a digestion of mathematics dealing with the concept of parsimony is given as well as how the concepts and equations apply to multivariate calibration in chemometrics. A procedure is then presented for selecting the most parsimonious model structure for a given calibration problem.

THEORY OF THE PARSIMONY PRINCIPLE

The concept of parsimony can be considered in a limited statistical context in multiple linear regression problems with variable selection. For a regression problem

$$\mathbf{c} = \mathbf{R}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{c}$ is a vector containing the physical or chemical property of interest (e.g. concentration) for a series of calibration samples, $\mathbf{R}$ is the response matrix containing λ measurements for each of the calibration samples, $\mathbf{b}$ is a vector of regression parameters and $\mathbf{e}$ is a vector of errors. The mean squared error (MSE) of the estimated concentration $\hat{c}$ can be written

$$\begin{aligned}\mathrm{MSE}(\hat{c}) &= E(\mathbf{R}\beta - \mathbf{R}\mathbf{b})^T(\mathbf{R}\beta - \mathbf{R}\mathbf{b}) \\ &= E[\beta - E(\mathbf{b})]^T\mathbf{R}^T\mathbf{R}[\beta - E(\mathbf{b})] \\ &\quad + E[E(\mathbf{b}) - \mathbf{b}]^T\mathbf{R}^T\mathbf{R}[E(\mathbf{b}) - \mathbf{b}]\end{aligned} \tag{2}$$

where E denotes the expected value over estimates of the model and β is the true model. Eqn. 2 shows that the MSE of $\hat{c}$ can be written as a bias squared (first term) plus a variance (second term). As parameters are added, or additional variables are included in the model, the bias of $\hat{c}$ will decrease while the variance of $\hat{c}$ will increase. $\mathrm{Var}(\hat{c})$ increases because the variance of the $\mathbf{b}$ increases as more parameters are estimated [10–13]. Therefore, variables should be added to the model only as long as the bias decreases more than the variance increases. In statistics then, the concept of parsimony arises when choosing the number of variables which minimize the MSE.

Several other references to the parsimony principle have been made in the literature, dating back as far as the 14th century to Occam's Razor (*Entia paraeter necesitatem non sunt multiplicanda*) [14]. A formal mathematical treatment which is more rigorous and generally applied than the discussion above was given by Stoica and Söderström [14] and Gustavsson et al. [15]. The application given in these papers is that of process identification, and therefore is time series data. This, however, does not limit the applicability of the theory. A summary of their work is presented below, and then the theory and assumptions as applied to multivariate calibration are discussed.

The mathematics in the rest of this section prove the statement "If two meaningful model structures are hierarchical or nested (i.e., one structure can be obtained by constraining in some way the other structure), and if the parameter estimation method is statistically efficient, then, for a fairly general class of validation criteria, the simpler structure is asymptotically better." [14].

Let c be a physical or chemical property of interest for a sample and $\mathbf{r}$ be be a vector containing the analytical measurements made on that sample. The true relationship can be described by a stochastic system L,

$$\mathrm{L}: \quad c = E\{c\,|\mathbf{r}\} + \epsilon \tag{3}$$

where E denotes the expected value of c given $\mathbf{r}$ and ϵ is the true measurement error. c may be a scalar or a vector. It is assumed that the ϵ are uncorrelated over all samples and that L holds for all samples under consideration. A model for L is introduced, denoted by $\mathrm{M}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is a finite dimensional vector containing the model parameters. That is,

$$\mathrm{M}(\hat{\boldsymbol{\theta}}): \quad c = f[\mathbf{r}, \hat{\boldsymbol{\theta}}] + e_{\mathrm{M}}(\hat{\boldsymbol{\theta}}) \tag{4}$$

where $e_{\mathrm{M}}(\hat{\boldsymbol{\theta}})$ are the residuals, and f is some function that is differentiable with respect to an admissible set of parameters. As the parameters vary over this admissible set, Eqn. 4 describes a set of acceptable models, called a model struc-

ture, denoted by M. Assume M is identifiable, that is, there exists a unique parameter vector $\boldsymbol{\theta}^*$ in M such that the model residuals equal the true measurement errors,

$$e_M(\boldsymbol{\theta}^*) = \epsilon \tag{5}$$

for all samples. The reasons for developing a model may be for example to determine certain parameters which have some specific meaning, for control system design, or for prediction of future sample properties. In chemometrics, models are most often used for prediction. No matter what the reason, the adequacy of the model must be evaluated via a 'validity function'. This criterion $V_M(\hat{\boldsymbol{\theta}})$ must be a scalar function of the parameter vector $\hat{\boldsymbol{\theta}}$ such that

$$V_M(\boldsymbol{\theta}^*) = \min_{\substack{\text{admissible} \\ \hat{\boldsymbol{\theta}} \text{ in M}}} V_M(\hat{\boldsymbol{\theta}}) \tag{6}$$

For example, V could be prediction error, calculated by root mean square error of prediction (RMSEP). Notice that $V_M(\hat{\boldsymbol{\theta}})$ evaluates the adequacy of a model with particular parameter values $\hat{\boldsymbol{\theta}}$. The interest here, however, is in evaluating the validity of the overall model structure M. Therefore, the expected value of V over all $\hat{\boldsymbol{\theta}}$ in M needs to be calculated. In order to do this, $V_M(\hat{\boldsymbol{\theta}})$ is expressed in a Taylor series expansion about $\boldsymbol{\theta}^*$. This gives

$$\begin{aligned} V_M(\hat{\boldsymbol{\theta}}) = V_M(\boldsymbol{\theta}^*) &+ \frac{dV_M^T}{d\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &+ \tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T V_M''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &+ \text{higher order terms} \end{aligned} \tag{7}$$

where

$$\frac{dV_M}{d\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) = \left[\frac{\partial V_M}{\partial \hat{\boldsymbol{\theta}}_1} \cdots \frac{\partial V_M}{\partial \hat{\boldsymbol{\theta}}_p}\right]\Bigg|_{\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*}$$

and p = the number of parameters in $\boldsymbol{\theta}$. $(dV_M/d\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^*)$ is a vector of zeros by Eqn. 6. Also by definition,

$$V_M''(\boldsymbol{\theta}^*) = \begin{bmatrix} V_{11} & \cdots & V_{1p} \\ \cdots & \cdots & \cdots \\ V_{p1} & \cdots & V_{pp} \end{bmatrix}_{\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*}$$

where

$$V_{ij} = \frac{\partial^2 V_M}{\partial \hat{\boldsymbol{\theta}}_i \partial \hat{\boldsymbol{\theta}}_j}(\boldsymbol{\theta}^*) \tag{8}$$

Assuming the higher order terms to be small, the expected value of Eqn. 7 can be approximated by

$$\begin{aligned} E\left[V_M(\hat{\boldsymbol{\theta}})\right] \approx V_M(\boldsymbol{\theta}^*) \\ + \tfrac{1}{2}E\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T V_M''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right] \end{aligned} \tag{9}$$

Since $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{V}_M''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is a scalar, Eqn. 9 can be written as

$$\begin{aligned} E\{V_M(\hat{\boldsymbol{\theta}})\} \approx V_M(\boldsymbol{\theta}^*) + \tfrac{1}{2}E\Big\{\text{tr}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \\ \times V_M''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\Big]\Big\} \end{aligned} \tag{10a}$$

$$\begin{aligned} \approx V_M(\boldsymbol{\theta}^*) + \tfrac{1}{2}E\Big\{\text{tr}\Big[V_M''(\boldsymbol{\theta}^*) \\ \times (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\Big]\Big\} \end{aligned} \tag{10b}$$

where tr indicates matrix trace. The only term in Eqn. 10b which depends on $\hat{\boldsymbol{\theta}}$ is $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T$. For unbiased parameter estimates, the expected value of this expression is the covariance matrix of the estimates. If the $\mathbf{P}_M = \text{cov}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, then Eqn. 10b can be written

$$E\left[V_M(\hat{\boldsymbol{\theta}})\right] \approx V_M(\boldsymbol{\theta}^*) + \frac{1}{2N}\text{tr}\left[V_M''(\boldsymbol{\theta}^*)\mathbf{P}_M\right] \tag{11}$$

where N is the number of times the parameters are estimated. If $V_M(\boldsymbol{\theta}^*)$ is constant over different model structures, it is sufficient to look at

$$W_M = \text{tr}\left[V_M''(\boldsymbol{\theta}^*)\mathbf{P}_M\right] \tag{12}$$

when comparing model structures. W_M is always greater than or equal to zero since $V_M''(\boldsymbol{\theta}^*)$ is positive definite and $\mathbf{P}_M$ is non-negative definite [16]. $V_M''(\boldsymbol{\theta}^*)$ is positive definite since it is a Hessian matrix evaluated at a minimum point. $\mathbf{P}_M$ is non-negative definite because it is a covariance matrix [17]. W_M equals zero only when $V_M''(\boldsymbol{\theta}^*)\mathbf{P}_M$ is the zero matrix, which occurs if the expected value of all $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{\theta}^*$. Eqn. 11 would then show the

expected value of the validity function to be $V_M(\theta^*)$. It would be exact in fact, since $(\hat{\theta} - \theta^*)$ would be zero. Therefore, a small value of W_M indicates a good model structure with respect to the validity function, V. Note further that W_M is a function of the parameter estimation method through $\mathbf{P}_M$, and the chosen validity function through $V''_M(\theta^*)$.

Thus far the discussion has concentrated on the properties of one model structure M. In order to prove the parsimony principle two model structures must be compared. Consider two model structures which fulfill

$$L \in M_1 \subset M_2 \tag{13}$$

L is the true functional relationship being sought, and the M_i are sets of model parameters or regions in a parameter space. Here the parameter vectors in M_2 have more elements than the parameter vectors in M_1. A further constraint for M_1 to be a subset of M_2 ($M_1 \subset M_2$) is that the model structure M_1 must be obtained by constraining some of the parameters in $\hat{\theta} \in M_2$ to be 0. Then these model structures are called 'hierarchical'. For example, if

$$M_1: \quad c = \theta_1 r_1 + e \tag{14a}$$

$$M_2: \quad c = \theta_1 r_1 + \theta_2 r_2 + e \tag{14b}$$

then M_1 and M_2 are hierarchical because M_1 is obtained by setting $\theta_2 = 0$. In contrast, if

$$M_1: \quad c = \theta_1 r_1 + \theta_2 r_2 + e \tag{15a}$$

$$M_2: \quad c = \theta_2 r_2 + \theta_3 r_3 + e \tag{15b}$$

then M_1 and M_2 are not hierarchical. A graphical interpretation of Eqns. 14a and 14b is shown in Fig. 1. Note further in Fig. 1 that the true model L is an element in M_1. For Eqn. 15 the parameter space is three dimensional; M_1 and M_2 each span a plane, and the two intersect along the $\theta_1 = \theta_3 = 0$ line. Having L be an element of M_1 and M_2 assures that $V_M(\theta^*)$ in Eqn. 11 is constant over M_1 and M_2.

Assume further that the parameter estimation method is 'statistically efficient' [18,19]. Then, for any validity function V which satisfies Eqn. 6,

$$W_{M_1} \leq W_{M_2} \tag{16}$$

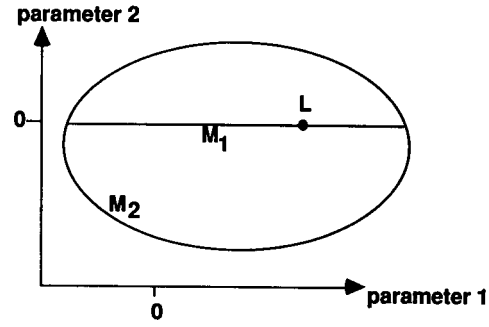


Fig. 1. Sets M_1 and M_2 in parameter space. M_2 has two parameters which vary inside the solid curve, and M_1 has one parameter which varies along the straight line. Here M_1 can be obtained by setting parameter 2 equal to zero. Notice $L \in M_1$.

This means that of two meaningful model structures, the one that is described by a fewer number of parameters will, on average, have a smaller validity function value asymptotically, or as the number of calibration samples gets large, proving the parsimony principle stated in the beginning of this section. The reason this is an average property is because the expected value of the validity function was taken over all acceptable realizations of the model parameters (Eqn. 9). The asymptotic property results from using an asymptotic property of statistically efficient estimation methods. The complete proof can be found in Ref. 16, but is also outlined in Refs. 14 and 15.

To review, several key assumptions made this proof possible. First, it is assumed that the model structures were hierarchical. Second, the validity function is a scalar function of the model parameters reaching a minimum at θ^*. Third, it is assumed that the true functional relationship, L, is contained in M_1. This means that in order to invoke the parsimony principle in practical applications, some information about L must be available or assumed. Finally, the parameter estimation method is assumed to have an asymptotic property of statistically efficient methods. In the next section each of the assumptions is addressed with respect to multivariate calibration. Also, the admissible set of parameter vectors for multivariate calibration is defined.

DISCUSSION

In this section the assumptions and features of the parsimony principle will be discussed with respect to multivariate calibration.

Hierarchical model structures

Multivariate calibration methods in chemometrics can be classified into linear and non-linear methods, and have been discussed extensively in the literature [4,5]. Details pertinent to establishing hierarchy are included below.

Linear inverse calibration model structures have the general form

$$c = \mathbf{r}^T\mathbf{b} + b_0 + \epsilon \tag{17}$$

where c is the physical or chemical property of interest, $\mathbf{r}$ is the response vector containing λ measurements, $\mathbf{b}$ is a vector of regression parameters and b_0 is the intercept. Similarly, the non-linear methods considered here satisfy

$$c = \mathbf{r}^T\mathbf{b} + f(\mathbf{r}) + b_0 + \epsilon \tag{18}$$

where f is some (continuous or piecewise) non-linear function of $\mathbf{r}$. Note this model structure still has a linear component. As defined above, the linear and non-linear model structures are hierarchical; $M_{linear} \subset M_{non\text{-}linear}$ as long as the non-linear model structure contains a linear term. That is, if $f(\mathbf{r}) = 0$, the non-linear structure reduces to the linear structure. Below the linear and non-linear model structures are considered in more detail, in order to further define the order of hierarchy.

Linear model structures. The linear model structures can be broken into two groups according to whether any measured variables are removed from the model before the model parameters are estimated. The method called ILS (inverse least squares) obeys the following linear model structure:

$$c = b_0 + b_1 r_1 + \ldots + b_k r_k + \epsilon \tag{19}$$

with $r_1 \ldots r_k$ selected from $\mathbf{r}$, and $k < \lambda$. The optimum k variables can be selected for example in a stepwise procedure [20,21]. Two other variable selection methods are described in Ref. 9 and 14. The model parameters b_i are then estimated by least squares. A second linear model where variables are selected is as follows: instead of selecting variables from $\mathbf{r}$, variables are selected from a rank reduced representation of the calibration data. This is essentially like ILS, except the variables are noise filtered or smoothed variables, and this will give biased estimates of the parameters. Here bias is accepted as a trade-off for lowering the variance of the parameter estimates. For this model structure the number of parameters is $k + 1$, the number of variables selected plus one for the intercept. The number of parameters is listed in Table 1, along with the number of parameters for the other model structures to be discussed.

TABLE 1

Number of parameters for each of the calibration methods

Method	Number of parameters [a]
ILS	$k+1$
PCR, PLS	$\lambda+1$
LWR	$\lambda+2$
PNL	$\lambda+1+g(\lambda)$ [b]
NN	$1+\lambda+h\lambda+h+h=1+\lambda+(\lambda+2)h$

[a] k = number of variables selected; λ = number of variables measured; r = pseudorank; h = number of hidden nodes; $r(\lambda - 1)$ is the number of parameters used to estimate the loadings for r factors. [b] g is related to the non-linear function. If using a biased estimate, $g(\lambda) > (\lambda - 1)r$.

The model parameters can also be estimated without removing any variables. Again, they can be estimated using least squares given sufficient conditions (if the number of calibration samples is greater than λ and the calibration matrix has full column rank), or biased least squares. Using a biased estimate leads to for example the familiar PCR and PLS, where a least squares estimate of the model parameters is made with a rank reduced (that is biased) representation of the calibration data. See Table 1 for the number of parameters in this model structure. If $k < \lambda$, $M_{ILS} \subset M_{linear}$; the model structures with variable selection are a subset of the model structures with no variable selection. This is seen by setting the model parameters equal to zero for the $\lambda - k$ variables in the linear structure to obtain the ILS model structure.

Non-linear model structures. Three classes of non-linear model structures are considered below; locally weighted regression, parametric and non-parametric modeling. Note that each of these structures has a linear component, and so follows the general form of Eqn. 18.

Locally weighted regression is a weighted linear regression method that has been implemented successfully for non-linear data [22–24]. It is based on the assumption that a smooth non-linear function can be locally approximated by a line. Since the model structure is piecewise linear over the range of concentration values, it cannot be considered a linear model structure. The weighted regression is accomplished by multiplying the calibration concentrations and responses by a diagonal weight matrix, **W**,

$$\mathbf{Wc} = \mathbf{WRb} + \mathbf{1}b_0 + \epsilon \tag{20}$$

where **1** is a vector of ones. **W** is a function of one parameter, the fraction of points near the unknown sample. Also needed to calculate **W** is the unknown sample response but this does not add to the number of parameters in the LWR model structure. Therefore, the number of parameters in a LWR model structure is $\lambda + 2$; λ for the regression vector, 1 for the intercept, and 1 for the fraction of points used. The regression parameters can be estimated using least squares where appropriate, or biased least squares. What is obtained for each prediction is a linear model valid over a small range of concentration values. The non-linear function f in Eqn. 18 is therefore a piecewise function describing deviations from some linear model. If the weight matrix is the identity matrix, LWR will be a purely linear model structure. Therefore, $M_{linear} \subset M_{LWR}$ since the linear model structure can be obtained by setting the weight function to have zero dependence on the fraction parameter, giving

$$M_{ILS} \subset M_{linear} \subset M_{LWR} \tag{21}$$

It can be shown that any smooth function can be fit to an arbitrarily small error with piecewise linear segments [25]. This means that ideally $L \in M_{LWR}$ for all L. However, the distance over which any particular line segment will be applicable may be very small; perhaps even smaller than the spacing of the calibration data points. So, $L \in M_{LWR}$ is not guaranteed unless the calibration points are infinitely dense. This implies that the more calibration samples there are that evenly span the calibration space, the better the chance of having $L \in M_{LWR}$. This is consistent with the suggestions for experimental design given in the literature where this method is introduced [22–24].

In parametric non-linear modeling a non-linear relationship between **r** and c is linearized by taking non-linear functions of **r** to form additional variables. A model which is then linear in the parameters is postulated. This can be considered to be parametric non-linear modeling (PNL) since a non-linearity is explicitly postulated. As with the strictly linear structure, the model parameters can be estimated using least squares or biased least squares methods. Non-linear PCR and non-linear PLS are two methods which use biased estimation; non-linear functions are taken of the scores from a PCA or PLS decomposition [23,26–30]. Unlike in the linear structure, the number of parameters in the model changes with the parameter estimation method. The reason for this is that with the biased estimator the loadings from a PCA or PLS decomposition are used explicitly. Consider a calibration model structure with linear and squared terms using biased estimation,

$$c = b_0 + \mathbf{r}^T\mathbf{b}_1 + (\mathbf{r}^T\mathbf{V})^2\mathbf{b}_2 + \epsilon \tag{22}$$

where the notation $(\mathbf{x})^2$ indicates squaring each element in the vector **x**. In this case the model has $1 + \lambda + r + (\lambda - 1)r$ parameters: b_0 is the intercept, $\mathbf{b}_1$ has λ parameters, $\mathbf{b}_2$ has r parameters, and **V** is a loadings matrix which has $(\lambda - 1)r$ parameters, where r is the pseudorank. The same model estimated by least squares methods

$$c = b_0 + \mathbf{r}^T\mathbf{b}_1 + (\mathbf{r}^T)^2\mathbf{b}_2 + \epsilon \tag{23}$$

has $1 + \lambda + \lambda = 1 + 2\lambda$ parameters. In general then, the number of parameters for parametric non-linear models will be $1 + \lambda + g(\lambda)$, where $g(\lambda)$ is a function of the non-linearity used and whether a least squares or biased least squares method is used to estimate the parameters (Table

1). With $g(\lambda) > 1$, M_{LWR} has fewer parameters than M_{PNL}. In order to establish hierarchy however, a LWR model structure must be obtained by setting some parameters in the PNL structure equal to zero. This cannot be accomplished, and so all that can be concluded is

$$M_{ILS} \subset M_{linear} \subset M_{PNL} \quad (24)$$

An important note is that in order for both LWR and PNL to remain hierarchical with M_{linear} no variable selection can be done.

The non-parametric non-linear methods to be discussed here are neural networks, projection pursuit regression and multivariate adaptive regression splines. The neural network architecture is one non-parametric modeling technique with a great deal of flexibility [31]. Considered here is a neural network structure with three layers: an input, middle (or hidden) and an output layer. The input layer has λ nodes, the middle layer has h nodes, and the output layer has 1 node. This neural network structure also has a linear part, relating the inputs (responses) linearly to the output (c) [2,32]. This structure is shown graphically in Fig. 2, with $\lambda = 3$, $h = 1$. In Fig. 2 each open circle (node) indicates a scalar value, and each line indicates a weight used to scale the node value. The output from each hidden node

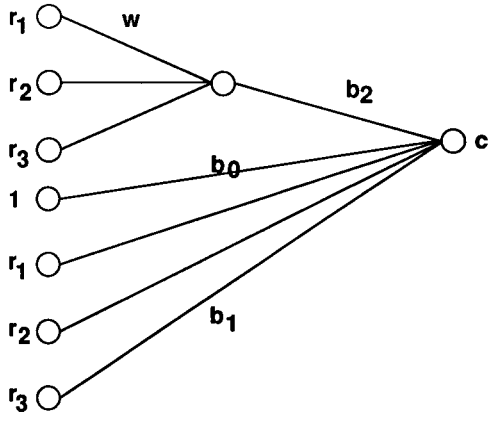


Fig. 2. Three-layer neural network model structure with direct linear feedthrough; $\lambda = 3$, $h = 1$. $\mathbf{b}_1$ is a vector containing the weights connecting all the responses linearly to the output c.

h, $o_h(\mathbf{r})$, is typically a sigmoid with one free parameter, α,

$$o_h(\mathbf{r}) = \frac{1}{1 + e^{-(\mathbf{w}_h^T \mathbf{r} + \alpha_h)}} \quad (25)$$

where $\mathbf{w}_h$ is a vector of weights that need to be estimated. The general model structure then is

$$c = b_0 + \mathbf{r}^T \mathbf{b}_1 + \sum_{i=1}^{h} \mathbf{b}_{2i} o_i(\mathbf{r}) + \epsilon \quad (26)$$

When $h = 0$ this neural network (NN) structure falls to a strictly linear model structure. The model parameters are typically estimated by minimizing a least squares criterion. The number of parameters for the structure in Eqn. 26 is $1 + \lambda + \lambda h + h + h$ (Table 1). The parameters may also be estimated via biased methods. This essentially reduces to minimizing a least squares criterion using scores as inputs instead of the raw data [32]. The projection of the measurements to a lower dimensional space is a linear transformation, and so can be collapsed into the transformation from the input to the hidden layer. Therefore, it does not change the number of parameters in the model, but may serve to increase the numerical stability of the parameter estimation method [32]. Neural networks are not hierarchical with PNL, since the PNL model structure cannot be obtained by setting parameters in M_{NN} equal to zero. Depending on the choices of h and the non-linear function in the PNL structure, M_{NN} may in fact have more or less parameters than M_{PNL}. M_{NN} is not hierarchical with M_{LWR} either. However, NN structures with different values of h are hierarchical; $M_{NN_1} \subset M_{NN_2}$ where NN_i means $h = i$. Therefore,

$$M_{ILS} \subset M_{linear} \subset M_{NN_1} \subset \ldots \subset M_{NN_h} \subset \ldots M_{NN_\infty} \quad (27)$$

It can be shown that for a three layer structure as described here, $L \in M_{NN}$ for all L [33]. That is, this neural network structure can model any function. However, to achieve a desired level of accuracy, h and the number of calibration samples may need to be very large. Two other non-parametric modeling methods are projection pursuit regression (PPR) and multivariate adaptive re-

gression splines (MARS). These are discussed in detail in Refs. 1, 3, 5, 34 and 35. Both of these methods give a linear model if the data are linear, and so are hierarchical with linear model structures. As with NN, within any one technique the model structures are hierarchical as the complexity is increased. PPR has the following model structure:

$$\mathbf{c} = \sum_{j=1}^{n} S_i(\mathbf{R}\alpha_i) + \epsilon \qquad (28)$$

where S_i is a smoothing function, and α_i is a vector indicating a linear combination of variables from **R**. A smoothing function is a piecewise linear function where each local segment spans the same range along $\mathbf{R}\alpha_i$. Eqn. 28 can also be written

$$\mathbf{c} = \sum_{i=1}^{n} \mathbf{R}\gamma_i + \epsilon \qquad (29)$$

where the γ_i are a function of the α_i and the smooth, S_i. If the γ_i are made to have zero dependence on the smoothing function S_i, PPR reduces to a linear model structure.

MARS is similar to PPR in that it uses a sum of piecewise functions to relate the responses to the concentrations. Here a relationship is drawn between the concentrations and individual responses from **R** (as opposed to a linear combination as in PPR) using function estimates which are either piecewise linear or piecewise cubic. In contrast to a smoothing function, the span over which any segment is valid is variable, defined by the adaptive 'knot' locations. MARS has an additional feature which allows variable interaction terms ($\mathbf{r}_i{}^*\mathbf{r}_j$) to be calculated. In both PPR and MARS it is assumed that there is a small number of independent variables (< 20). This may not be the case for example in spectroscopic applications. Recently it has been suggested that in order to be able to more generally use these methods, the principal component or PLS scores should be used rather than the raw data [35]. This essentially gives biased versions of these methods.

By constraining MARS to use piecewise linear functions with evenly spaced knot locations and zero interaction terms, a PPR model structure is obtained. Since these methods are similar to LWR in that local models in response space are constructed,

$$\mathrm{M_{ILS}} \subset \mathrm{M_{linear}} \subset \mathrm{M_{LWR}} \subset \mathrm{M_{PPR}} \subset \mathrm{M_{MARS}} \qquad (30)$$

The hierarchy of the calibration methods discussed above is summarized graphically in Fig. 3 where the arrows ($\rightarrow$) indicate subset ($\subset$) relationships.

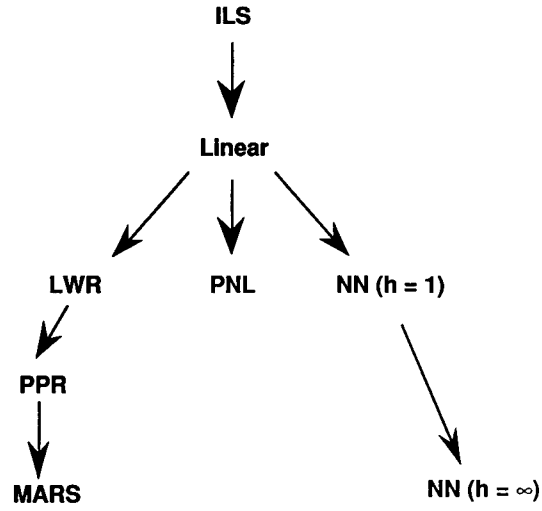


Fig. 3. Graphical summary of the hierarchy structure of multivariate calibration methods. The arrows ($\rightarrow$) indicate subset ($\subset$) relationships.

Validity function

As was mentioned, it is often of interest in multivariate calibration to minimize the prediction error, or more formally, $E\{[c - \hat{c}(\hat{\boldsymbol{\theta}})]^2\}$, where the expected value is taken over all future samples. For the application of the parsimony principle to multivariate calibration the validity function can therefore be written as

$$V_{\mathrm{M}}(\hat{\boldsymbol{\theta}}) = E\left\{\left[c - \hat{c}(\hat{\boldsymbol{\theta}})\right]^2\right\} \qquad (31)$$

This function will reach a minimum when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$, because here the model errors will match exactly the concentration errors, meaning no imbedded error has been included in the calculation of the model parameters. This choice of validity function is therefore acceptable for this theory.

Admissible parameters

As was discussed in the Theory section, when the parameter vector varies over an admissible set, Eqn. 4 describes a family of models, referred to as model structure. The admissible set is a small subset of all the possible parameter vectors (which is all of $\mathscr{R}^p$ where p is the number of model parameters). If the parameter vectors were not restricted, the expected value of the validity function would be large for every model structure. This is because nearly all of the $\hat{\theta} \in \mathscr{R}^p$ give rise to unrealistic models with large validity function values. The admissible set then is made of parameter vectors that are in some way appropriate. A definition for admissible parameter vectors for multivariate calibration is given below.

An admissible model is one that is modeling the data well, yet not overfitting. Proposed here is that an admissible $\hat{\theta}$ fits three criteria:

(1) The parameters are estimated using some statistical method, i.e. estimated by least squares, not just randomly selected from $\mathscr{R}^p$; (2) Require the $e(\hat{\theta})$ to have the same distribution as ϵ and (3)

$$\tau - \delta \leq \mathrm{var}\left[e(\hat{\theta})\right] \leq \tau + \delta \qquad (32)$$

where $\tau = \mathrm{O}[\mathrm{var}(\epsilon)]$ and δ is chosen so that for all admissible $\hat{\theta}$, $\mathrm{var}[e(\hat{\theta})]$ and τ are not statistically different given the available degrees of freedom.

The first criterion assures that the parameter vectors are in some way reasonable. The second criterion requires that the model has no systematic errors past what the noise properties of the calibration model dictate. If, for example, the ϵ are randomly distributed over the concentration range, $e(\hat{\theta})$ must also be random when plotted as a function of concentration. Example $e(\hat{\theta})$ are shown in Fig. 4. If the true errors are random over the concentration range, the model giving rise to Fig. 4a would be admissible while the $\hat{\theta}$ leading to Fig. 4b would not. One assumption in least squares is that the model errors are random, and independent of the concentration. If the true errors are not of this form, a weighted least squares model must be made (Ref. 20, Sections 5.2 and 5.3). There are several ways of evaluating

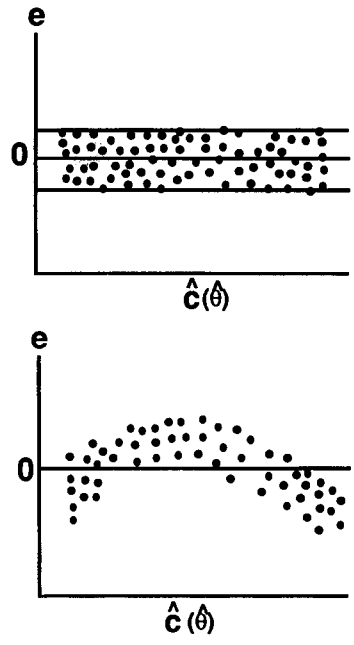


Fig. 4. Example concentration residuals $[e = c - \hat{c}(\hat{\theta})]$ plotted as a function of the estimated concentration. (a) Random residuals, with a variance of τ (denoted by the horizontal lines), (b) non-random residuals.

the distribution of the errors as a function of **c**, including using autocorrelation of the residuals [36].

The purpose of the third criterion is to require that the concentration residuals are as small as can be expected based on the errors in the calibration data, yet not smaller, as this is an indication of overfitting. That is, overfitting occurs when errors from the calibration data are included in the model parameters, and the fit error is smaller than the prediction error. When the uncertainty in the analytical measurements is small compared to the laboratory uncertainty of the concentrations, $\tau \approx \mathrm{var}(\epsilon)$. It is not well understood how the $\mathrm{var}[e(\hat{\theta})]$ scales up when the measurement errors are large. One derivation is given by Lorber and Kowalski [37], but is applicable only for linear model structures. τ may be externally defined as well, according to the requirements of the application. The δ is included to allow the variances $e(\hat{\theta})$ to lie in an interval where the values are not statistically different from τ given a chosen confidence interval.

By definition, θ^* fits these criteria. Since the model errors match the true errors (Eqn. 5), the distributions must be equal. Also, var[$e(\theta^*)$] will be in the specified interval since it equals var(ϵ).

A non-empty model structure M means that there exists at least one $\hat{\theta} \in M$ such that the above properties hold. Below a procedure is given to find the most parsimonious non-empty M for a given calibration application.

$L \in M_1$

In the theory section, $L \in M_1$ was assumed to be true. In practice such an assumption is not easily made. However, the rather restrictive definition used for admissible parameter vectors essentially assures that L is an element of any non-empty M to the certainty that the data will permit. Noise in the calibration data will not affect the location of L, but it does affect how close a model estimated with noisy data can get to L. Therefore, there is a region around L in parameter space wherein the models cannot be differentiated from L. An interesting result is that L may really lie in a non-linear region, but due to this uncertainty, may overlap into the linear region. Therefore, even if the true L is non-linear, an adequate model may be found using a linear model structure. This is shown graphically in Fig. 5.

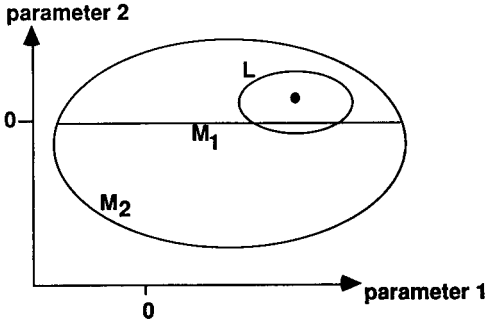


Fig. 5. Plots of the M_{linear} (M_1) and $M_{non\text{-}linear}$ (M_2) in a two-dimensional parameter space. Here L is the point where the true model is located, but due to noise in the calibration data, has a region around it in which the models cannot be differentiated from L. Note that in this case the region extends into the linear model structure domain.

Parameter estimation

In the discussion of theory, the parameter estimation methods were assumed to be unbiased (Eqns. 10 to 11) and statistically efficient. In particular, an asymptotic property of statistically efficient methods was used in the proof of the parsimony principle. Asymptotic means that as the number of calibration samples, I, approaches infinity. The estimation methods used in multivariate calibration are least squares and biased least squares. Least squares is unbiased, and is statistically efficient [18,19]. Below is an explanation for why the biased regression methods will asymptotically be unbiased and efficient.

As has been pointed out, the biased methods can be expressed in two steps [38]:

(1) form a few new variables by a linear combination of the original variables (which often results in data smoothing);

(2) least squares estimation using the new variables.

For example, PCR can be interpreted as least squares using the first r score vectors from a principal component decomposition. The score vectors are the coordinates of the calibration samples along a new set of coordinate axes. The relationship of the new coordinates to the original coordinates (the measurements) is contained in the loadings.

The least squares model assumes the variables are independent. In methods where PCA and PLS are used to obtain the new variables, independence is guaranteed since the new variables are orthogonal. The question that remains is if the new variables are biased from the true variables.

Consider now estimating the new variables as $I \to \infty$. As I gets large, the significant loadings will approach the true loadings; that is, the loadings calculated from noise free data. The scores on these significant axes will reflect the true signal along that axis, and so admissible parameters estimated (by least squares) with these variables will be unbiased. In contrast, if I is small, the loadings will be biased from the true loadings. See for example Fig. 6 which shows the first loading for a simulated data set calculated with no noise (solid) and also with noise and finite

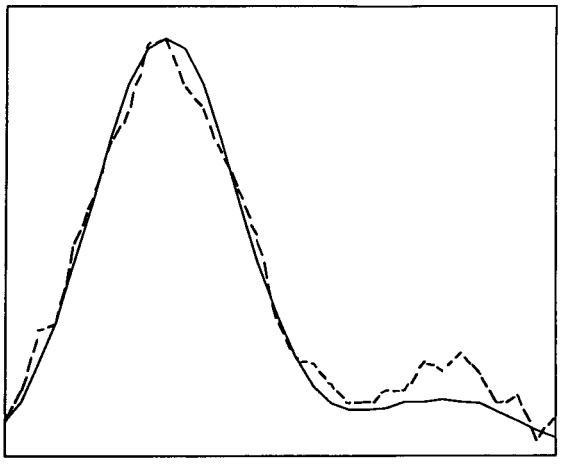


Fig. 6. The first principal component loading for a data set that is noise free (solid) and has measurement error (dashed). The measurement axis may be for instance a wavelength axis.

samples (dashed). The scores on these axes will contain therefore the truth plus bias, and so admissible parameters estimated with these new variables will be biased. In multivariate calibration the bias is accepted as a tradeoff for lowering the variance of the parameter estimates. Therefore, methods that use new variables estimated by a linear combination of the original variables to estimate admissible parameter vectors will be asymptotically unbiased and statistically efficient. This means that the assumptions regarding the parameter estimation method hold, and the parsimony principle can be applied in multivariate calibration.

Use of the parsimony principle for choosing a parsimonious model structure

The theory of the parsimony principle can now be applied to select the most parsimonious calibration model given only a calibration set of data. Presented here is a sequence of model structures to be considered. By the parsimony principle, once a non-empty model structure M has been found using the criteria specified above, it will be preferred over any structure it is a subset of with respect to minimizing the validity function, V. However, not all multivariate calibration methods are hierarchical, and the parsimony principle cannot be invoked in all circumstances.

The objective is to find the most parsimonious non-empty model structure, and so the sequence of methods will follow the order of hierarchy, shown in Fig. 3. The ILS model structure is the most parsimonious structure considered here. If admissible parameter estimates are found as defined in the *Admissible parameters* section, all structures that are hierarchical with it do not need to be considered according to the parsimony principle. All calibration methods discussed in this paper are hierarchical with ILS.

If an admissible model cannot be found with the ILS structure, a linear model structure with all the variables should be considered next. Within this structure, several methods are available, including PCR and PLS. If an acceptable model can be found, all structures that are hierarchical with it do not need to be considered. Except for ILS, all the calibration structures discussed here are hierarchical with M_{linear}.

Given that an admissible linear model cannot be found, a non-linear model structure must be considered. However, the non-linear structures are not all hierarchical, and so it is not clear in what order the model structures should be considered (that is, down which branch to proceed in Fig. 3). The different structures have different numbers of parameters, and this can be used to guide which branch to follow first. The statements that can be made however, will not be as strong since the parsimony principle cannot be universally invoked beyond the branch point.

The non-linear model structure with the fewest number of parameters is LWR. The only reason an acceptable model would not be found with this structure is if the calibration samples do not densely cover the calibration space. If admissible model parameters are found using this structure, three statements can be made:

(1) by parsimony principle this model structure is preferred over the PPR and MARS model structures;

(2) this model structure may or may not be preferred over PNL or NN. Since they are not hierarchical the PP cannot be used;

(3) $L \in M_{LWR}$ according to the assumption made in the $L \in M$ section.

The remaining model structures, PNL, PPR, MARS and NN have different numbers of parameters depending on the exact model structure that is postulated. Therefore, no clear guidelines can be given for a sequence of method to try. Chemical understanding of the application can help to guide the user. For instance, if the functional relationship is expected to change over the concentration range, a local modeling procedure should be used. Similarly, if the functionality is expected by theory to be the same non-linear function over the concentration range, perhaps using PNL or NN should be the next step. Also not that the parsimony principle can be used within a model structure, i.e., $NN_1 \subset NN_2$, etc.

The model structures presented here all have a linear component. If in fact L is strictly non-linear the most parsimonious model structure may be very simple, for example $\mathbf{c} = \mathbf{r}_1^2$ which is not hierarchical with any structure in Fig. 3. Diagnostics are needed for providing the user with a means for choosing a model structure with no linear term. Methods such as PPR and MARS may be helpful in these cases as they can provide graphical estimations of linearizing transformations.

Conclusions

The issue addressed in this paper is how to select a best calibration method for a given application. It is not a trivial problem, since there are now a number of methods from which to choose. Given the single goal of minimizing prediction error, it was shown that of two meaningful models which are hierarchical, the one with fewer parameters is preferred. A procedure was presented which can be used to find the most parsimonious model for a given application. Note however, that a simple model may not have adequate outlier detection capabilities. In fact, optimal outlier detection (using probably many variables) is critical to the long-term reliability of the value predicted by the parsimonious model.

MBS acknowledges the financial assistance of The Calgon Corporation for a Merck Fellowship. This work was also supported by the Center for Process Analytical Chemistry (CPAC), a National Science Foundation/Industry/University Cooperative Research Center at the University of Washington. Special thanks are due to Age Smilde and Sonja Sekulic for their contributions while at the University of Washington.

REFERENCES

1 J. Friedman and W. Stuetzle, J. Am. Stat. Assoc., 76 (1981) 812.
2 J. Long, V. Gregoriou and P. Gemperline, Anal. Chem., 62 (1990) 1791.
3 J. Friedman, Ann. Stat., 19 (1991) 1.
4 B.R. Kowalski and M.B. Seasholtz, J. Chemom., 5 (1991) 129.
5 S. Sekulic and B.R. Kowalski, J. Chemom., 6 (1992) 199.
6 M. Stone, J. Roy. Stat. Soc. B, 36 (1974) 111.
7 B. Efron and G. Gong, Am. Stat., 37 (1983) 36.
8 M.B. Seasholtz, D.A. Archibald, A. Lorber and B.R. Kowalski, Appl. Spectrosc., 43 (1989) 1067.
9 P.J. Brown, J. Chemom., 6 (1992) 151.
10 R.R. Hocking, Biometrics, 32 (1976) 1.
11 L. Breiman and D. Freedman, J. Am. Stat. Assoc., 78 (1983) 131.
12 J. Friedman and B. Silverman, Technometrics, 31 (1989) 3.
13 A. Höskuldsson, Chemom. Intell. Lab. Sys., 14 (1992) 139.
14 P. Stoica and T. Söderström, Int. J. Control, 36 (1982) 409.
15 I. Gustavsson, L. Ljung and T. Söderström, Automatica, 13 (1977) 59.
16 T. Söderström, L. Ljung and I. Gustavsson, Lund Institute of Technology, Department of Automatic Control, Technical Report No. 7428 (1974).
17 G. Golub and C. Van Loan, Matrix Computations, Johns Hopkins, Baltimore, MD, 1989.
18 E. Greenberg and C. Webster, Advanced Econometrics: A Bridge to the Literature, Wiley, New York, 1983.
19 P. Eykhoff, System Identification, Wiley, New York, 1974, pp. 410–415.
20 N.R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 2nd edn., 1981, pp. 337–341.
21 W. Hrushka, Data Analysis: Wavelength Selection Techniques, in P. Williams and K. Norris (Eds.), Near Infrared Technology, American Institute of Cereal Chemists, St. Paul, MN, 1987.
22 W. Cleveland and S. Devlin, J. Am. Stat. Assoc., 83 (1988) 596.
23 T. Næs, T. Isaksson and B.R. Kowalski, Anal. Chem., 62 (1990) 664.
24 T. Næs and T. Isaksson,, Appl. Spectrosc., 46 (1992) 34.
25 M.J.D. Powell, Approximation Theory and Methods, Cambridge University Press, New York, 1981.

26 T. Isaksson and T. Næs, Appl. Spectrosc., 42 (1988) 1273.
27 N.B. Vogt and K. Kolset, Chemom. Intell. Lab. Syst., 6 (1989) 221.
28 N.B. Vogt, Chemom Intell. Lab. Syst., 7 (1989) 119.
29 S. Wold, N. Kettaneh-Wold and B. Skagerberg, Chemom. Intell. Lab. Syst., 7 (1989) 53.
30 I. Frank, Chemom. Intell. Lab. Syst., 8 (1990) 109.
31 P. Jansson, Anal. Chem., 63 (1991) 357A.
32 P. Gemperline, R. Long and V. Gregoriou, Anal. Chem., 63 (1992) 2313.
33 G. Cybenko, Approximations by Superpositions of a Sigmoidal Function, Math. Contr., Signals Syst., 2 (1989) 303.
34 K. Beebe and B.R. Kowalski, Anal. Chem., 60 (1988) 2273.
35 S. Sekulic and B.R. Kowalski, J. Chemom., submitted for publication.
36 G.E.P. Box and G.M. Jenkins, Time Series Analysis, Holden-Day, San Francisco, CA, 1976.
37 A. Lorber and B.R. Kowalski, J. Chemom., 2 (1988) 93.
38 H. Martens and T. Næs, Multivariate Calibration, Wiley, New York, 1989.

Analytica Chimica Acta, 277 (1993) 179–188
Elsevier Science Publishers B.V., Amsterdam

Colored information from a black box?

Validation and evaluation of neural networks

G. Kateman and J.R.M. Smits

Laboratory for Analytical Chemistry, Katholieke Universiteit Nijmegen, Toernooiveld 1, 6525 ED Nijmegen (Netherlands)

(Received 3rd September 1992; revised manuscript received 2nd December 1992)

Abstract

Artificial neural networks are used in many fields and have proven their worth. Until recently the applications in analytical chemistry are scarce, but now many groups are working in this field. One of the reasons for the hesitation in accepting artificial neural networks is the weak scientific understanding of the system. In fact it works as a black box. However, the applications in analytical chemistry so far are so encouraging that more research will be conducted. One very important aspect will be the critical validation of artificial neural networks as long as no sound theoretical background is available. Some criteria will be discussed and some practical examples will be given.

Keywords: Artificial neural networks; Black box; Neural networks

The principles of neural networks are quite well known nowadays. It may be sufficient to memorize the origin of this type of algorithms: the brain. In a brain cells, neurons, are connected with sensors, active sites and with each other. A number of actions performed on a single neuron result in a firing of an output action of this neuron. This network is able to the seemingly endless range of brain activities that are characteristic of brain activity: memorizing, logical inference, actions upon an impulse, pattern recognition, etc. A simple model can be implemented on a computer, in hardware or software. This model is always simpler than the brain because the number of artificial neurons that can be simulated always is much less than the number of neurons in a real brain. Moreover the not yet fully understood (chemical) dynamical process in the brain can be simulated only rather primitive. The simplest model of neural network is the one of Fig. 1: a number of input units is connected to output units. Every connection has a weight; each output neuron has a threshold value. When patterns, combinations of activated and non-activated input neurons are fed to the input, these activities are fed through the weight function to each of the output neurons. When these weighted activities are summed in each output neuron and the threshold value is exceeded, the output node gives off a signal. On this theme many variations are possible. A much used model has an extra layer of neurons, called a hidden layer, in between input and output layer. The weights can be adapted automatically with different algorithms, the threshold function can have many shapes and the connection pattern of the neurons can be very diverse.

It must be remarked that the connection between brain and artificial neural network is not liked by many researchers, be it because the brain

Correspondence to: G. Kateman, Laboratory for Analytical Chemistry, Katholieke Universiteit Nijmegen, Toernooiveld 1, 6525 ED Nijmegen (Netherlands).

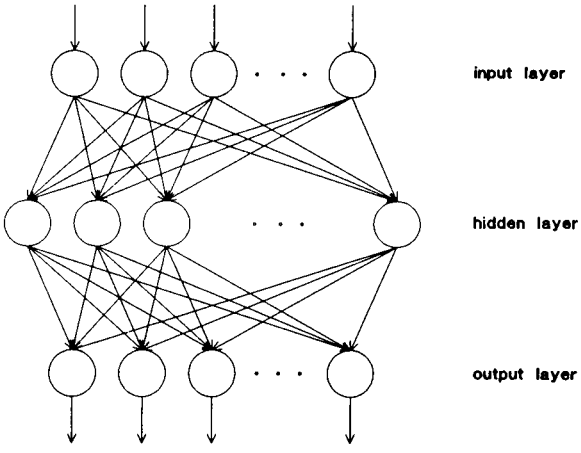


Fig. 1. Simple artificial neural network.

outperforms the artificial net by lengths, or because it limits the thinking about new configurations. So instead of neurons often the terms nodes or units are used. Neural networks are called a special type of parallel distributed systems, etc. In this paper we will stick to the terms connected with neurons.

As an example the configuration of a much used network and algorithm will be explained (see e.g., [1]).

MULTILAYER FEEDFORWARD NETWORKS

The choice of the network type depends on the problem that has to be solved. Multilayer feedforward networks [2] are most often used. In these networks signals are propagated from the input layer through the hidden layers(s) to the output layer. A neuron or unit thus receives signals via connections from other neurons (or the outside world). These connections are associated with weights, and the net input for a neuron j is given by

$$\mathrm{net}_j = \sum_i w_{ji} o_i$$

in which i represents neurons in the previous layer, w_{ji} is the weight associated with the connection from neuron i to neuron j, and o_i is the output of neuron i.

The output of a neuron is determined by the transfer function and the net input of the neuron. A popular transfer function is the sigmoid

$$o_j = \mathrm{f}(\mathrm{net}_j) = \frac{1}{1 + \mathrm{e}^{-(\mathrm{net}_j + \theta_j)}}$$

in which θ_j is a bias which shifts the function along the net_j axis.

BACKPROPAGATION LEARNING RULE

The adequate functioning of a neural network depends very much on the way the signals are propagated through the network. The weights play an important role in this propagation and a proper setting of these weight factors is essential. Generally, such a setting is not known beforehand and initially the weights are given small random values. The process of adapting the weights to an optimal set of values is called the training of the neural network. Usually, this training is done by means of supervised learning. A representative training set with examples (problems with their known solutions) is presented iteratively to the neural network and the difference between the desired and the obtained solution is used to adapt the weights in small steps, according to a learning rule.

A much used learning rule is called the backpropagation learning rule [2]. The adaption of the weights is given by

$$\Delta w_{ji}(n+1) = \eta \delta_j o_i + \alpha \Delta w_{ji}(n)$$

in which η is the learning rate and α the momentum term. If the learning rate is low, the convergence of the weights to their optimum value is very slow and one may get stuck in a local optimum. If it is too high, the system may oscillate. To limit the danger of oscillation, the momentum term is used. δ_j depends on the layer. If layer j is the output layer, δ_j is given by

$$\delta_j = (d_j - o_j)\mathrm{f}_j'(\mathrm{net}_j)$$

in which d_j and o_j are the desired output and the obtained output of neuron j, respectively. If the layer is a hidden layer, δ_j is given by

$$\delta_j = \mathrm{f}_j'(\mathrm{net}_j)\Sigma \delta_k w_{kj}$$

in which k represents the neurons in the next layer. Usually some noise of the same level as the data noise will be added, as this prevents the algorithm to stay fixed in a sub-optimum.

Kohonen networks

Many other types of networks exist that differ more or less from the network described before. A different approach is the so-called Kohonen network [3]. This self-organizing network belongs to the class of topology preserving mapping techniques. This implies that during the learning phase the spatial ordering of the neurons in the map is driven by the topology, as present in the training set. As an example only two-dimensional Kohonen networks will be considered.

Training patterns are represented as n-dimensional vectors (arrays of variables). The network consists of a formal input array and an output matrix of neutrons. Each neuron in the output matrix possesses a weight vector which forms the connection between this particular neuron and all neurons in the input array. After training, application of an input vector produces different neuron output values at different positions in the output matrix (the so-called Kohonen map). The neuron with the weight vector closest to the input vector is assigned to represent that object. Basically, the Kohonen learning algorithm can be divided into 6 stages [4,5]:

(1) Assign for each neuron j in the network the components of its weight vector $\vec{w}_j$ to random values.

(2) Select pseudo-randomly an input vector $\vec{x}$.

(3) Determine for each neuron j the distance D_j between its weight vector $\vec{w}_j$ and the input vector $\vec{x}$. Using Euclidian metrics, the distance measure can be expressed as:

$$D_j = \sqrt{\sum_{i=1}^{N} (x_i - w_{ij})^2}$$

where N denotes the total number of components in the vector $\vec{x}$ and w_{ij} is the weight of the connection between the ith component of the input $\vec{x}$ and neuron j.

(4) Assign the neuron having a weight vector as close as possible to $\vec{x}$ (i.e., D_i is at minimum) as the winner. In addition, select all neurons in close vicinity of the best matching neuron too.

(5) Adapt the weight vectors of the selected set of neurons according to:

$$\vec{w}_j^{(new)} = \vec{w}_j^{(old)} + \eta\left(\vec{x} - \vec{w}_j^{(old)}\right)$$

where η represents the learning rate.

(6) Repeat steps 2–5 until a sufficient degree of convergence has been achieved.

The weight vectors of a winning neuron and its neighbors are gradually rotated towards the applied input vector. By this mechanism, similar features present in the input pattern set are mapped onto the same region of the two-dimensional map. However, there is no prior knowledge which neuron in the map will be associated with a specific input pattern or class of input patterns. Furthermore, despite of the preservation of topology, in case a high dimensional space is mapped onto just two dimensions information on the real distance between input vectors mapped on two nonadjacent neurons in the network is lost in great deal. Several strategies concerning the adaptation of the learning rate and the size of the neighboring region have been proposed in literature [2,6]. In the system as described by Melssen and Smits [5] for instance the following rules of thumb were adopted:

(1) The learning rate η was decreased linearly with the number of iteration steps.

(2) A neuron was assigned to be a neighbor of the winning neuron if it was located within a square-shaped region centred around the winning neuron. The width of this square was stepwise decreased after a predefined number of iterations.

POSSIBILITIES

As the artificial neural network tries to mimic the brain in essence, the possibilities of the network are comparable to the activities of the brain. The Kohonen network is suited to mapping, a network that compares its output with its input is a memory system, able to correct defects in some input patterns. The back-propagation network is

good in pattern recognition and generalizing, some special networks are suited to interpretation or extrapolation (prediction).

According to Zupan and Gasteiger [7] in analytical chemistry the input objects that have to be translated in network readable input patterns of neural-network studies can be roughly divided into four groups:

(i) Spectra for determining spectra/structure correlations.

(ii) Sequences of amino acids for determining the secondary structure of proteins.

(iii) Sets of signals from different sensors for diagnosis of malfunctions.

(iv) Other objects.

For special purposes many deviating networks have been developed. Peterson [8] describes a "counter-propagation" network. A potential advantage of counter-propagation networks over back-propagation networks is the absence of an adjustable learning parameter in its transfer function. One of the hidden layers utilizes an output function which allows for competition among the neurons of the layer. The weights change according to the Kohonen learning rule. In counter-propagation no adjustment of the learning coefficient was required.

Other variations are possible by changing the connection pattern as will be shown later.

CONNECTION BETWEEN ARTIFICIAL NEURAL NETWORKS AND MATHEMATICAL TECHNIQUES

Sources of nonlinear response in spectra are: curvature in the concentration–response function; shifts in the position of the absorption bands; and changes in the width of absorption bands.

The back-propagation training rule for neural networks is equivalent to the steepest descent technique of function minimization. The steepest descent technique is used to minimize the network output error by adjusting the weights to decrease the output error in the direction of a local downhill gradient. This technique performs poorly when the input variables are highly correlated. In order to improve the training speed and decrease the overall calibration error in neural networks, orthogonal transformations of the measured response variables can be used. Gemperline [9] uses the singular value decomposition, giving k principal components. The number of orthogonal variables, k, used in the network is selected by "pruning" input nodes from the network having variables with small singular values. Both the partial least squares (PLS) method and neural networks are soft modeling techniques. Data are separated in combinations of factors. When there are less factors than variables, less relevant information is removed. Principal components consist of scores and loadings, the loadings are comparable to the weights of the neural net.

In PLS the **X** and **Y** datamatrices are decomposed into principal components, sequentially for each principal component. The correlation between **X** and **Y** can be shown by fitting the scores with a linear function. When a polynomial regression is used, a somewhat better solution can be obtained. The problem is that the polynomial depends on the problem at hand. A neural net is always nonlinear and gives as a rule better solution for nonlinear problems.

For auto-association with linear output neurons [10], the optimal weight values can be derived by standard linear algebra, consisting essentially in singular value decomposition and making thus the nonlinear functions at the hidden layer completely unnecessary. This is similar in spirit to the Karhunen Loève transform.

ARCHITECTURE

The connections between the neurons or nodes can be made in an almost infinite way. Many attempts have been made to design schemes that are optimal for a special task.

In general one object entity (or variable) corresponds to one input neuron. However, Aoyama et al. [11] recently compared the prediction performance of two networks: one with six input variables and the other with the same six input variables plus six new inputs consisting of the squares of the first six.

Zupan and Gasteiger [7] suggest that in spite

of the demand for more input neurons, a distributed coding scheme should be tried more often. Besides QSAR, another area of potential use could be the spectra/structure correlation problem, especially for NMR and MS, where feeding several input neurons from one spectral feature might enhance the overall performance of the network. This speculation is based on the common spectroscopic knowledge that a specific spectral feature can receive contributions from several different structural entities. Hence it would seem beneficial if more input and hidden neurons received contributions from the same input feature.

Some form of experimental design technique should be applied. Most authors seem not to have given sufficient attention to this aspect. Selecting the right size and right examples for the training set should be the most carefully planned preliminary procedure of a neural network approach. According to Zupan and Gasteiger [7] most of the authors input all of the data they could lay their hands on into the neural nets. Others however believe that many weights allow better generalizing. In our opinion the number of patterns used should be chosen carefully, in the sense that the pattern set should give a balanced representation of the problem space. The number of inputs should be kept to a minimum.

Borggaard and Thodberg [12] describe a network with direct connections from input to output and without the sigmoid transformation in the output layer, the network becomes a true generalization of multivariate linear regression: if the connections from the hidden neurons are set to zero, the network reduces to MLR. In some cases only hidden neurons are needed, in others only direct connections, and in others both. When both types of connections are allowed in the same network, the task of deciding what is needed is left to the training algorithm.

Smits et al. [13] described flat versus modular neural-network systems for the interpretation of IR spectra. If division in frequency intervals is chosen as a representation of the spectra and fragment coding is chosen as a representation of molecules, two different approaches may be taken in applying neural network systems for interpreting IR spectra, namely a flat neural network system [14] or a modular one.

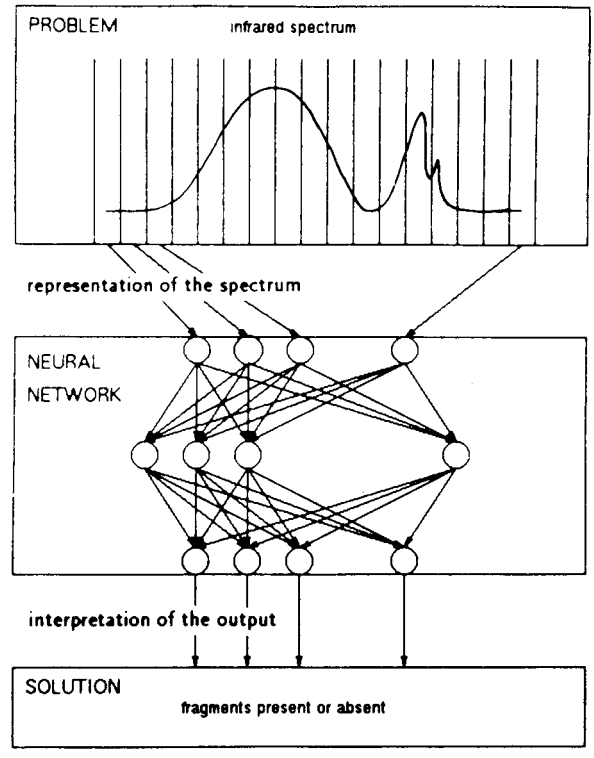


Fig. 2. Network using an IR spectrum as input.

A flat neural network system consists of one large neural network with on the input side the IR spectrum and on the output side all the possible functional groups, such as, e.g., in Fig. 2.

There are several advantages of this approach.

(i) The network takes all the functional groups into account at the same time, so that interferences may be perceived.

(ii) No prior knowledge other than the division into different functional groups is needed.

However, there are also some disadvantages.

(i) Functional groups which are difficult to learn in combination with other groups may decrease the performance of the entire system.

(ii) Different functional groups may require different optimal spectrum representations (broader or narrower intervals at different wavelength positions).

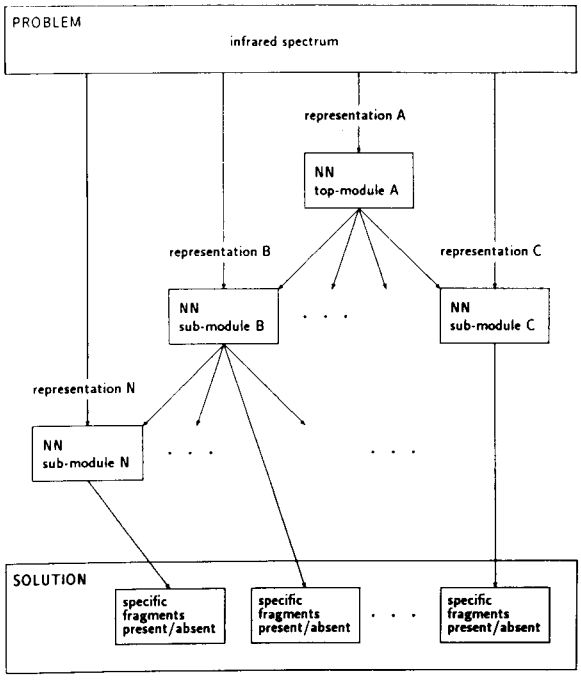


Fig. 3. Modular neural network.

(iii) Extension of the system to include additional functional groups requires much effort (unless no hidden layers are involved [14]).

(iv) Large networks are more difficult to optimize.

A modular neural network system consists of a number of small neural network modules, connected with each other in a tree-like structure (Fig. 3). Each network has its own specific task. The top-module also has the spectrum as input, but the output is only a rough division into sub-problems (classes). For each class a separate neural network module exists, specialized on a specific sub-problem. Each module has the spectrum as input (possibly represented differently), and a refinement of the sub-problem for which it is designed as the output. Further layers of modules may be used for further refinement, until the desired level is reached.

Advantages of this approach are:

(i) For functional groups that are difficult to learn in combination with other groups, modules may be designed to facilitate separate learning.

(ii) For different functional groups different spectrum representations may be used, so that the emphasis may be put on different parts of the spectrum.

(iii) Extension of the system to include other functional groups is relatively easy.

(iv) The networks are easier to optimize.

(v) Training sets can be balanced easier.

Disadvantages may be that:

(i) Correlations/interferences of functional groups belonging to different sub-problems are not perceived. This disadvantage may be partly neutralized by information exchange between the modules. The output of "rougher" networks may be used as additional input, next to the spectrum, for the "finer" networks.

(ii) The division into sub-problems is not straightforward and requires prior knowledge. According to Melssen et al. [5] the Kohonen network can provide this knowledge.

Not only the architecture of the networks can be changed, all parameters in the algorithm can be modified. Bos et al. [15] adopted the approach to introduce a local learning rate for each neuron which is the inverse of the number of inputs which the neuron has.

If the patterns cover the pattern space poorly, back-propagation produces large errors for outlier patterns. Weight adaptations are made for patterns making an error larger than ϵ. If none of the patterns produces an error larger than ϵ, it is lowered. With very long training ϵ will become low enough to use all patterns.

With a binary classification scheme, it is possible for a network to misclassify a sample by a number of ways. By turning the wrong output neuron "ON", a sample could be misclassified as is the case when none of the neurons is turned "ON". If more than one output neuron is turned "ON" an ambiguous classification occurs. With this in mind, Long et al. [16] monitored the progress of a network during the training process in terms of percent hits with the number of possible hits being equal to the number of data set samples multiplied by the number of classes. The learning rate and the momentum term were varied for each of the networks. Rubner et al. [17,18] developed self organizing networks. Fig-

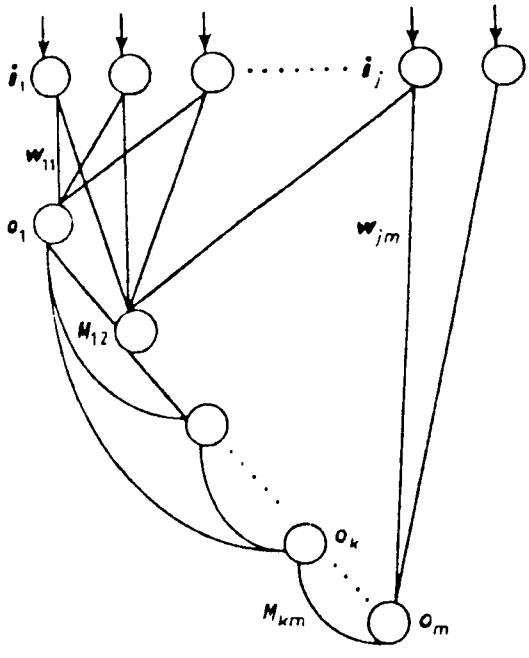


Fig. 4. Rubner's self organizing network.

ure 4 shows a scheme of the network. He proposes that these lateral weights adapt themselves according to an anti-Hebbian rule: the change of the lateral weight u_{im} between two output neurons l and m is negatively proportional to the product of pre- and post-neuronal activities. The weight vectors w_m converge to the eigenvectors c_α of the covariance matrix **C**. As a result, the output neurons correspond to analyzers of mutually orthogonal features that extract the directions of diminishing variance of the input patterns.

Jokinen [19] describes a dynamically capacity allocating network that is able to learn incrementally as more information becomes available and to avoid the spatially unselective forgetting of commonly used learning algorithms for neural networks. The method has many tuning parameters which have to be chosen by the designer. Unfortunately, the methods to select the parameters are based on heuristics which do not have a solid theoretical foundation and quite a lot of experience is needed to select the parameters properly.

REPRESENTATION

The representation of a problem (e.g., an IR spectrum) and its solution (a molecule) is not straighforward. Upon translation of a problem and a solution to their representations, as much information as possible has to be retained.

The problem representation is fed to the input neurons of the neural network (one character per neuron). The signal is propagated through the network to the output neurons, which give the solution representation. In case of a classification each output neuron represents a class and when one neuron has a high output value while the others have low output values the given pattern belongs to the class corresponding to that neuron.

When representing a spectrum another way has to be followed. An IR spectrum may be divided into intervals along the wavelength axis. The scaled mean absorbance/transmittance in each interval may then be used as value for each input neuron. A choice has to be made about the width of each interval. Making the intervals too narrow may lead to huge networks and the introduction of noise. Too broad intervals, on the other hand, may lead to a loss of information. When the neural network system is used purely as a classification system, a description of the chemical compound class is sufficient. Each output neuron represents a class of chemical compounds. When the neural network is used as an aid for interpreting a spectrum, a molecule can be represented by fragment coding. For IR spectroscopy commonly accepted fragments (functional groups) may be chosen. The advantage is the simplicity of the representation. Each output neuron is associated with a functional group and its value indicates the presence or absence of that specific functional group. A disadvantage, however, is that information about the structure of the molecule as a whole is lost, which may be of great importance, because of the interaction of different functional groups close together in the molecular structure. Melssen et al. [5] proposed using a Kohonen network for this purpose. An alternative seems to be a connection table in which relative positions of the different functional groups in the molecule are retained. Major

drawbacks of these tables are their large and variable size and the difficulty of assigning values to the input neurons for different structures. For these reasons connection tables appear to be of limited practical use.

It may be clear that neural networks are still in a phase of its development that numerous adaptions and configurations are possible and are used. Although there is a strong connection between mathematical methods (principal component analysis, singular value decomposition, nonlinear PLS, etc) the exact relation is not yet theoretically described. The question is whether this is a prerequisite for using neural networks for solving problems in analytical chemistry. In the first place analytical chemistry always used techniques that were useful for solving problems. When a good scientific base was available, the better, but when the method was purely empirical analytical chemists sought an other way. Validation and evaluation have always been an important tool in making the results of analytical chemical measurements acceptable. In the second place the role of statistics in providing a base for analytical chemical results cannot be overestimated, but a few remarks can be made. The use of chemometrical techniques under the heading "multivariate analysis" learns that these methods should be used with as much care as empirical chemical or physical methods and that these techniques should be scrutinized by the same validation and evaluation methods [20]. Comparing these mathematical techniques with neural networks may not be possibe on a strict mathematical base, but the results can be compared and are in many cases better for the networks, especially for nonlinear problems [21–23]. A problem is that the networks, as has been described before, can be changed in so many ways that it is nearly impossible to select a suitable network for a certain problem that has been evaluated thoroughly. Using formal optimizing methods, experimental designs or sequential methods, solves only a part of the problem, as the networks work statistically by design. Zupan and Gasteiger [7] suggested to discard the idea of presenting the patterns randomly and using a strict design. Others presented the idea of self organizing.

For the time being we are confronted with the problem of colored solutions. The colored glasses used by the experimenter allow him/her to use a network that suits his purposes well and has the danger that solution has companions that could be better suited to other circumstances. As long as there does not exist a theoretical coupling between neural networks and mathematical methods, and there is no rigorous evaluation of these mathematical methods we should rely on the time proven method of validation and evaluation.

VALIDATION, EVALUATION AND PERFORMANCE

Validation is defined as the test on whether the system responds to the requirements specified by the developer. Evaluation is defined as the test on whether the system corresponds to the expectations and requirements of the users of the system.

Validation

Several propositions are made on how to perform these tests. An often used validation test is the error of the training set or preferably the test set, usually expressed as RMS (root mean square), or NSE (normalized standard error). The NSE is given by:

$$\mathrm{NSE} = \frac{1}{2p} \sum_{p} \sum_{j} (d_{pj} - o_{pj})^2$$

in which d_{pj} and o_{pj} are the desired output and the actual output respectively. p represents the number of patterns in the data set and j represents the output neurons. An extension is the incorporation of the NSE of a test set. This test set can be chosen in many ways. One possibility is to take a part of the available data set by choosing randomly a number of patterns and declare this set the test set or prediction set. The size of the test set is arbitrary, a test set of one pattern is possible but then the process should be repeated until all patterns have been used as test set (leave-one-out method). The errors of training set and prediction set are monitored and training is

stopped as soon as no further improvement of the prediction set is observed.

If the resulting errors are of the same magnitude, one can assume that the underlying features of the pattern are learned and not only the specific features in the training set. Another possibility is to express this performance in the percentage of correct solutions.

Often the performance of a classification system is expressed in two percentages: recognition and prediction. The recognition and the prediction are both the number of correctly classified patterns of a set, divided by the total number of patterns present in this set. In case of the recognition the set of patterns is the same set as used for training the classification system, in case of the prediction the patterns in the set were not used for training. When the performance meets the conditions set in advance, the network may be used for real analysis.

Smits and co-workers [21,22] compared the results of a validated network with real data in the classification of algae.

Gemperline et al. [9] used a random presentation order for training patterns. Test data sets were analyzed at frequent intervals (ca. 20–50 iterations) to evaluate SEP (standard error of prediction). Training was stopped when SEP converged to an average minimum value.

$$\%\mathrm{SEP}_j = \frac{100}{\bar{c}_j}\left[\frac{\sum_{i=1}^{n}\left(\hat{c}_{ij} - c_{ij}\right)^2}{p}\right]^{1/2}$$

where $\hat{c}_{ij}$ is the estimated concentration of the jth component for the ith "unknown" and p is the number of "unknowns".

Bos and Weber [24] defined fitness as:

$$\text{fitness} = \left[\sum_{j=1}^{k}\sum_{i=1}^{l}(\text{required output} - \text{calc. output})^2\right]^{-1} = \frac{2}{\mathrm{NSE}}$$

where k is the number of samples and l the external outputs of the network. Smits et al. [13] tested the capabilities of a neural network by comparing the results with the performance of a genuine expert. The network consisted of a set of modules, each responsible for the classification of IR spectra according to a functional group in the sample. 47 spectra were selected. The selection of the spectra was biased towards cases for which the interpretation of the network was incorrect. Three questions were formulated: represents the spectrum an alcohol and/or carbonyl group, and within each group what specific group. The spectra were divided according to the questions and both the expert and the network were asked to answer them. A correct answer yielded 1 point, an incorrect answer no points and a question mark 0.5 points. The overall performance of expert and network was comparable although the relative performance depends somewhat on the problem. In all the expert gained 182 points, the network 190 from a maximum score of 220 points. The expert had 16 completely correct answers, the network 22 from 47 problems.

Evaluation

Borggaard and Thodberg [12] suggested to test for both generalization and robustness. The generalization is the performance on examples drawn from the same population as the training data, while the robustness is the performance on examples drawn from a somewhat different, but related population. The latter is pertinent for real-life applications, since things often are not as they were in the calibration procedure. He used two training sets: C, the calibration set and M, the monitoring set. For evaluation of the trained network he uses three sets: the T-set for testing the interpolation, drawn from the same population as C and M, the E1-set to test extrapolation in the variable to be determined and the E2-set to test extrapolation in another variable hampering the determination. E1 and E2 are drawn from populations from a somewhat different but related population.

Conclusion

Many parameters influence the performance of a neural network: purpose, architecture, momentum, threshold, learning rate, bias, added

noise, etc. Some configurations produce results that are more or less comparable to the results of statistical analysis, other configurations produce results that are in such a format that no comparison is possible, or able to provide solutions that in no other way can be obtained. Validation and evaluation are time-proven tools to obtain an insight in the reliability of the results. However, though statistical tests are often used in the validation, the protocol that describes the testing procedures is lacking usually. When neural networks are to gain confidence in chemometrics or analytical chemistry, a well described and thoroughly tested validation and evaluation procedure is indispensable. Maybe the work of Borggaard for estimations and the work of Smits for pattern recognition should be a starting point.

REFERENCES

1 R.P. Lippmann, IEEE ASSP Mag., April (1987) 4.
2 D.E. Rumelhart and J.L. McClelland (PDP Research Group), Parallel Distributed Processing, MIT Press, Cambridge, MA, 1986.
3 T. Kohonen, Self-Organization and Associative Memory, Springer Verlag, Heidelberg, 1984.
4 P.D. Wasserman, Neural Computing. Theory and Practice, Van Nostrand Reinhold, New York, 1989.
5 W.J. Melssen, J.R.M. Smits, G.H. Rolf and G. Kateman, Chemom. Intell. Lab. Syst., 18 (1993) 195.
6 H. Ritter and K. Schulten, Biol. Cybern., 60 (1988) 59.
7 J. Zupan and J. Gasteiger, Anal. Chim. Acta, 248 (1991) 1.
8 K.L. Peterson, Anal. Chem., 64 (1992) 379.
9 P.J. Gemperline, J.R. Long and V.G. Gregoriou, Anal. Chem., 63 (1991) 2313.
10 H. Bourlard and Y. Kamp, Biol. Cybern., 59 (1988) 291.
11 T. Aoyama, Y. Suzuki and H. Ichikawa, J. Med. Chem., 33 (1990) 2583.
12 C. Borggaard and H.H. Thodberg, Anal. Chem., 64 (1992) 545.
13 J.R.M. Smits, P. Schoenmakers, A. Stehmann, F. Sijstermans and G. Kateman, Chemom. Intell. Lab. Syst., 18 (1993) 27.
14 E.W. Robb and M.E. Munk, Mikrochim. Acta, (1990) 131.
15 A. Bos, M. Bos and W.E. van der Linden, Anal. Chim. Acta, 256 (1992) 133.
16 J.R. Long, H.T. Mayfield, M.V. Henley and P.R. Kromann, Anal. Chem., 63 (1991) 1256.
17 J. Rubner and P. Tavan, Europhys. Lett., 10 (1989) 693.
18 J. Rubner and K. Schulten, Biol. Cybern., 62 (1990) 193.
19 P.A. Jokinen, Chemom. Intell. Lab. Syst., 12 (1991) 121.
20 M.J.P. Gerritsen, H. Tanis, B.G.M. Vandeginste and G. Kateman, Anal. Chem., 64 (1992) 2042.
21 H.W. Balfoort, J. Snoek, J.R.M. Smits, L.W. Breedveld, J.W. Hofstraat and J. Ringelberg, J. Plankton Res., 14 (1992) 575.
22 J.R.M. Smits, H.W. Balfoort, L.W. Breedveld, J. Snoek, J.W. Hofstraat, M.W.J. Derksen and G. Kateman, Anal. Chim. Acta, 258 (1992) 11.
23 A.P. de Weijer, L. Buydens, G. Kateman and H.M. Heuvel, Chemom. Intell. Lab. Syst., 16 (1992) 77.
24 M. Bos and H.T. Weber, Anal. Chim. Acta, 247 (1991) 97.

Analytica Chimica Acta, 277 (1993) 189–197
Elsevier Science Publishers B.V., Amsterdam

FLIN: Fuzzy Linear Interpolating Network

Peter de B. Harrington and Brian W. Pack

Center for Intelligent Chemical Instrumentation, Department of Chemistry, Clippinger Laboratories, Ohio University, Athens, OH 45701–2979 (USA)

(Received 3rd September 1992; revised manuscript received 9th December 1992)

Abstract

A fuzzy linear interpolating network (FLIN) has been developed that is a local processing neural network. Local processing advantageously furnishes a traceable mechanism of inference and a bounty of diagonistic information in the variable scores and observation loadings of the processing units. FLIN is a two layer network for which the first layer accomplishes data driven model selection and the second layer provides linear predictive models. A new method of training is presented that enhances the relations between unsupervised and supervised layers of this network. The advantages of FLIN are demonstrated with a spectrophotometic titration of litmus.

Keywords: Principal components analysis; Spectrophotometry; FLIN; Fuzzy linear interpolating network; Neural networks

Artificial neural networks are computer systems that simulate biological neural networks, many of which are computer programs that run on digital machines. Digital systems are free of processing errors. Biological neural networks average out the processing errors by distributing information throughout the network. Because artificial neural networks are good models of biological systems, they also distribute information throughout the network. This step is unnecessary if the neural network is implemented on digital hardware. Another detriment to distributed processing is that the mechanism of inference becomes hidden among the multitude of connections. Neural networks are considered to be black box classifiers because of this characteristic.

Neurons in artificial neural networks will be referred to as processing units. For digital computers, alternative neural networks have been developed that localize information to particular units [1,2]. Redundancy is not required to prevent processing errors, so each unit may perform a distinct processing task. Localized systems are minimal neural networks in that a minimum number of neurons is required for inference. The advantage of localized processing is that diagnostic information is furnished by the path through the network and variable loadings of the unit weight vectors.

Neural networks (NNs) have been applied to calibration in analytical chemistry [3]. The counterpropagation neural network is useful for building calibration models [4]. It was developed by Hecht-Nielsen [5], and may offer a hundredfold increase in training efficiency over the more popular backpropagation network [6]. The counterpropagation neural network is a two layer network that is a Kohonen self-organizing map in the front layer [7] and a Grossberg outstar in the second layer [8]. These two layers will be referred to as the K-layer and the G-layer, respectively.

Correspondence to: P. de B. Harrington, Center for Intelligent Chemical Instrumentation, Department of Chemistry, Clippinger Labs., Ohio University, Athens, OH 45701–2979 (USA).

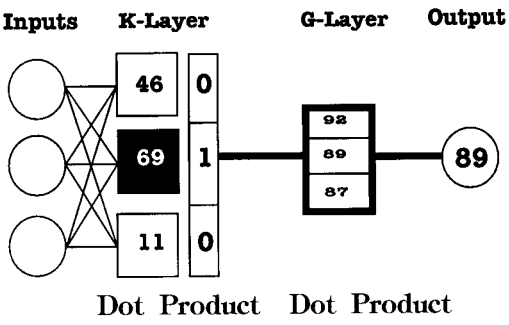


Fig. 1. Schematic of a counterpropagation network operating in normal mode.

Figure 1 gives a schematic of a simple counterpropagation network with three K-units and one G-unit.

Information in a neural network is stored by the synaptic strengths of connections between axons and dendrites of a neutron. The synaptic strength between the conections of input unit j and output unit i is represented by a weight w_{ij}. In a counterpropagation network all the units perform the same function.

$$o_i = \sum_{j=1}^{v} w_{i,j} x_j \qquad (1)$$

The number of inputs or units in the previous layer is v. The output of the ith neuron (o_i) is obtained from the dot product of the input ($\mathbf{x}$) and the weight ($\mathbf{w}_i$) vectors.

The units are trained differently in the two layers. K-layer training is unsupervised. Unsupervised learning extracts information soley from the network inputs, and is a type of cluster analysis [9]. The unit with the largest output is adjusted by moving it closer to the input.

$$^*w_{i,j} = w_{i,j} + \eta(x_j - w_{i,j}) \qquad (2)$$

The updated weight $^*w_{i,j}$ is obtained by adding a small increment (η) of the difference between the input and old weight value. This increment is the learning rate and typically ranges between 0.1 and 0.001. If η is large the network may not converge, and if η is small convergence may require a prohibitively long time. Kohonen training moves the weight vectors toward the center of clusters in the multidimensional data space. In the normal mode of counterpropagation the largest output is set to unity and all other outputs to zero.

Training the G-layer occurs after the K-layer has been trained. This method is supervised and is an example of delta learning [10]. Supervised learning extracts information by minimizing the difference between the training outputs and the network outputs.

$$^*w_{i,\max} = w_{i,\max} + \eta(o_i - t_i) \qquad (3)$$

The component of the weight vector that corresponds to the maximum output from the previous layer $^*w_{i,\max}$ is updated by multiplying the learning rate times the difference between the unit output (o_i) and the target value (i.e., actual output) t_i. Each unit in the G-layer corresponds to a network output variable. The weight vectors of the G-units function as lookup tables that are indexed by the K-layer for retrieving an output (see Fig. 1). The correlations between input and output are distributed throughout the network.

The fuzzy linear interpolating network (FLIN) is an adaptation of the counterpropagation neural network. FLIN uses localized processing and first order models for non-linear responses. FLIN is a neural network implementation of non-linear prediction by regional models [11,12]. Counterpropagation networks use zero order models in the G-layer. The system described is a prototype system that uses localized processing. This paper demonstrates a modification of a counterpropagation training algorithm that enhances the relationships between the two layers. The K-layer training has become more supervised and the G-layer has become less supervised.

THEORY

The FLIN network functions in a similar manner as the counterpropagation network. The first layer performs a model selection step by finding the unit which most closely resembles the input vector. This unit will direct the model in the G-layer that actually makes the prediction. Only a single unit in the G-layer will fire furnishing the outputs. The G-units will contain prediction ma-

trices instead of a single output value. Localized processing occurs at the G-unit making the prediction and the K-unit that selects the path. The input to the K-layer is processed and passed to the G-unit in the second layer. In distributed systems the unit outputs are passed to the following layer.

Training the FLIN network requires two modes of training. The first mode of training is unsupervised and is referred to as exploration mode. The units search for clusters in the data space and each cluster models the relationship between input and output variables. The clusters are obtained during training so that they will minimize distance (MD) between inputs and K-units. The second mode is supervised and is referred to as focus mode. In this mode clusters in the data space are expanded or shrunk, so that they minimize prediction errors (ME) (see Fig. 2A and B).

The K-layer selects a model by a Euclidean distance calculation from the center of gravity stored in the units. Instead of a dot product the distance is calculated from the input to the weight vector of each unit.

$$d_i = \frac{\sqrt{\sum_{j=1}^{v} (x_j - k_{i,j})^2}}{\sigma_i} \qquad (4)$$

The relative distance (d_i) from unit i and the v-dimensional input ($\mathbf{x}$) is calculated in Eqn. 4. The $\mathbf{k}_i$ measures the recognition or memory of the unit. Only a single unit wins and selects the unit G-layer for predicting the output. The radial size of the data cluster around $\mathbf{k}_i$ is obtained from a recursive measure of the standard deviation, σ_i. This parameter is set to a value of unity during exploration mode. The cluster size is obtained during focus mode and will be discussed later. The output from the unit with the minimum distance is the input vector with $\mathbf{k}_{md}$ subtracted from it.

$$^*x_j = x_j - k_{md,j} \qquad (5)$$

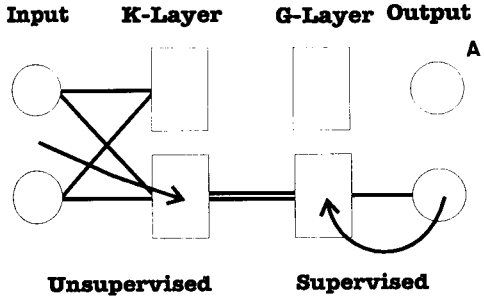
The input to the G-layer ($^*\mathbf{x}$) is output from the winning K-layer unit.

In exploration, mode $\mathbf{k}_i$ represents a cluster center in the input data space (see Fig. 2A). Each $\mathbf{k}_i$ is initialized to the average observation of the data set. During training the $\mathbf{k}_i$ with the shortest distance between itself and a training input $\mathbf{x}$ is adjusted by

$$k_{md,j} = k_{md,j} + \eta(x_j - k_{md,j}) + \mu \Delta k_{md,j} \qquad (6)$$

The training rate of $\mathbf{k}_{md}$ may be accelerated by the use of a momentum term [6,10]. This term adds the momentum factor μ times the previous adjustment $\Delta\mathbf{k}_{md}$ to the unit. This form of training is similar to that used by counterpropagation networks because K-layer training is unsupervised and G-layer training is supervised.

In training the K-layers a fuzzy algorithm is used that improves the modeling of the data space. A triangular fuzzy membership function is



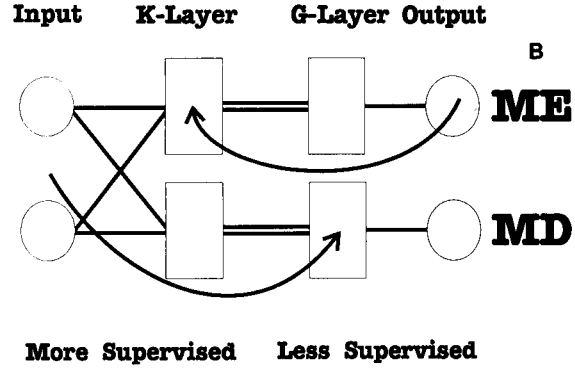


Fig. 2. (A) Schematic of a FLIN network operating in exploration mode. (B) Schematic of a FLIN network operating in focus mode. The units with minimum prediction error are indicated by ME and the untis with minimum distance are indicated by MD.

used for inputs that appear between the two closest K-vectors. When this occurs the projection of the input onto the difference vector between the two K-units should be greater than zero.

$$\sum_{j=1}^{v}(k_{md,j}-k_{m2,j})(x_j-k_{m2,j})>0$$
$$\sum_{j=1}^{v}(k_{m2,j}-k_{md,j})(x_j-k_{md,j})>0 \quad (7)$$

If the two conditions described by Eqn. 7 are satisfied, **x** is considered between the closest (md) and next closest (m2) K-units. The fuzzy membership value (f) is the multivariate projection onto the difference vector.

$$f=\frac{\sum_{j=1}^{v}(x_j-k_{m2,j})(k_{md,j}-k_{m2,j})}{\sum_{j=1}^{v}(k_{md,j}-k_{m2,j})^2} \quad (8)$$

This membership function is a linear function of the projection of **x** on the difference vector ($\mathbf{k}_{md}-\mathbf{k}_{m2}$). Both units are adjusted using the product of f and η for the adjustment in Eqn. 6.

The G-layer is also trained at the same time. These units perform a dot product calculation on the outputs from the K-layer.

$$\mathbf{o}_{md}=\mathbf{G}_{md}{}^{*}\mathbf{i}+\mathbf{b}_{md} \quad (9)$$

Matrix notation is used for Eqn. 9. The number of input variables is v and output variables is p. A vector of outputs $\mathbf{o}_{md}$ is obtained from the G-unit that is connected to the K-layer unit that is closest to the input. The output from the K-layer $^{*}\mathbf{i}$ is multiplied by a (pxv) weight matrix ($\mathbf{G}_{md}$). The unit also has a p-dimensional vector of bias values that are added to the matrix product.

The delta learning rule is applied to the G-units during training.

$$g_{md,j,k}=g_{md,j,k}+\eta^{*}x_j\times(o_k-t_k)+\mu\Delta g_{md,j,k} \quad (10)$$

The G-unit that is connected to the K-unit that fires has its predictor matrix $\mathbf{G}_{md}$ updated by the outer-product of the input vector ($^{*}\mathbf{x}$) and the difference between the network output (**o**) and the target output (**t**). A similar procedure is used with

$$b_{md,k}=b_{md,k}+\eta(o_k-t_k)+\mu\Delta b_{md,k} \quad (11)$$

the bias vector ($\mathbf{b}_{md}$). Both equations contain a momentum term which adds a factor μ of the previous weight adjustment ($\Delta g_{md,j,k}$,$\Delta b_{md,k}$).

The focus mode of training focuses the units by minimizing the prediction error in an indirect manner. This mode of training is less supervised than Grossberg training and more supervised than Kohonen training (see Fig. 2B). In this mode, the unit with the lowest prediction error (ME) is updated by Eqn. 2, and the K-layer unit with the shortest distance (MD) to the input has its predictor matrix updated. Additionally, the K-layer units start modeling the cluster size by calculating σ_i

$$\sigma_i=\alpha\sqrt{\sum_{j=1}^{v}(k_{me,j}-x_j)^2}+(1-\alpha)\sigma_i \quad (12)$$

The learning rate for cluster size is α. When K-units are connected lower errors will expand in size. As σ decreases in size for a K-unit, it may no longer be the closest unit to any of the training inputs. For this case, the K-unit will not be adjusted during training, and the unit is essentially removed from the network.

The FLIN network first is trained in exploratory mode. Each cycle through the training set is referred to as an epoch. When the error converges FLIN will shift into focus mode. The units will reorganize the clusters by varying the radial size in the K-layer and models in the G-layer to lower prediction error. The number of units in K and G layers is always equal and connected in a one to one manner.

Principal component analysis (PCA) was used to compress the data set, so that overdetermined (OD) and underdetermined data could be evaluated [13]. The principal components have two useful properties. They maximize variance and are orthogonal. If the number of principal components (p) is less than the number of variables, a data reduction may be obtained. An observa-

tion score ($\mathbf{r}_i$) is obtained by projecting the data matrix **D** onto a principal component ($\mathbf{c}_i$).

$$\mathbf{r}_i = \mathbf{D}\mathbf{c}_i \quad (13)$$

FLIN is trained on the observation scores. The original data space may be obtained by an inversion procedure.

$$\mathbf{D}^* = \sum_{i=1}^{p} \mathbf{r}_i \mathbf{C}_i^{\mathrm{T}} \quad (14)$$

The reconstructed data ($\mathbf{D}^*$) will approach the original data matrix as p is increased.

EXPERIMENTAL

Synthetic data set 1 was obtained by using a bivariate normal distribution with a 0.2 standard deviation around four cluster means; (−2, 1), (−1, −1), (1, −1), (2, 1). This data set was constructed, so that the advantage of fuzzy K-training could be observed.

All calculations were performed on a 33-MHz Insight 80486 computer which was equipped with 8 Mb of RAM and 128 Kb RAM cache. The code was developed and debugged using Borland C 3.1 and evaluated in 32 bit DOS extended mode using Watcom C 8.5 ver G.

Litmus indicator was prepared by weighing 1.75 g of litmus powder (Fisher Scientific Cat. No. L-140) and dissolving in 50 ml of distilled water and 50 ml ethanol. The solution was then heated to enhance dissolution and the supernatant was decanted and used in the experiment as the indicator. A simple strong acid–strong base titration was performed using HCl and NaOH. The strength of the acid was 0.1 N and was prepared from concentrated stock HCl (Mallinckrodt, Cat. No. 2612-6 × 6, Lot 3560 KBMV). The base used as the titrant was VWR Scientific NaOH (Lot A2-17, Cat. No. VW6744-4). The litmus indicator was added to the NaOH titrant and absorption spectra were taken at 0.5 ml intervals from 0 ml to 19.5 ml of titrant. The absorption spectra were then converted to ASCII XY format by Galactic Industries Grams/386. All visible spectra were obtained using a Hewlett Packard 8452A diode array spectrophotometer with a 1 cm × 1 cm plastic cell. Spectra contained 211 points and were collected at a 2 nm^{-1} data interval from 400 to 820 nm. Forty spectra were collected at a 0.5-ml titration interval. They were divided into evaluation and calibration data sets.

The principal components were calculated from the mean centered absorbance spectra using singular value decomposition [14]. The scores on the first two principal components were used. These components respectively accounted for 96.6 and 3.3% of the total variance and together virtually represented all the variance in the data set.

The FLIN training parameters used a momentum (μ) of 0.75. The cluster modeling parameter (α) was set to 1.0×10^{-2}. The average error for 100 epochs was used for parameter adjustment.

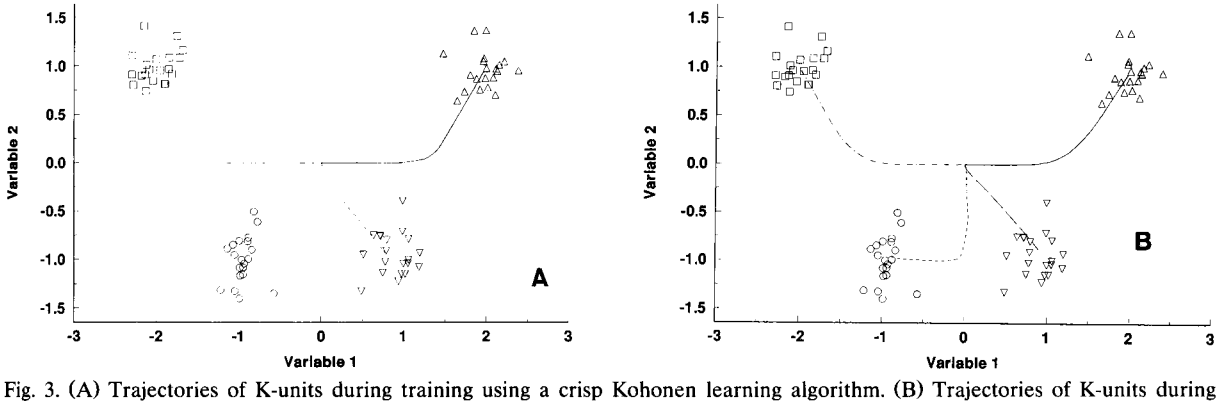


Fig. 3. (A) Trajectories of K-units during training using a crisp Kohonen learning algorithm. (B) Trajectories of K-units during training using a fuzzy Kohonen learning algorithm.

The learning rate (η) was initially set to 1.0×10^{-2}. When the change in error was less than 1.0×10^{-3}, η was reduced to 1.0×10^{-3}. Once η is reduced, when the change in error drops below 1.0×10^{-3}, FLIN is set in focus mode and η is increased back to its initial value of 1.0×10^{-2}. The learning rate is decreased similar to that used in the exploration mode when the change of error is below 1.0×10^{-3}. After this point, when the change in error drops below 1.0×10^{-3}, training is terminated.

Training times are given in real time for two unit networks. The PCA compressed data (40 observations and 2 variables) require 2.5 min which included the PCA calculation. The uncompressed data set (40 observations and 211 variables) required 25 min to train.

RESULTS AND DISCUSSION

The first evaluation demonstrates the need for fuzzy K-training. Fig. 3A and B gives the K-vector trajectories for the crisp and fuzzy learning algorithms. The trajectories plot the K-vectors at each epoch of training. Fuzzy learning is advantageous because units are allowed to share inputs during training thereby yielding more homogenous distributions of the K-vectors.

The second evaluation models the change in litmus color during a tritration. The concentration of an indicator is measured spectrophotometrically. Litmus will undergo a spectral transition at the end point of an acid-base titration which yields a bilinear model. For this experiment the indicator is added to the titrant, so that the number of moles of indicator is increasing during the titration and when the volume correction below is applied the concentration of indicator will be proportional to titration volume

$$*x_i = x_i \frac{\text{sample} + \text{titrant}}{\text{sample}}$$

The volume corrected absorbance of spectrum x at wavelength i is obtained by multiplying by the total volume divided by the sample volume. The endpoint is indicated in Fig. 4.

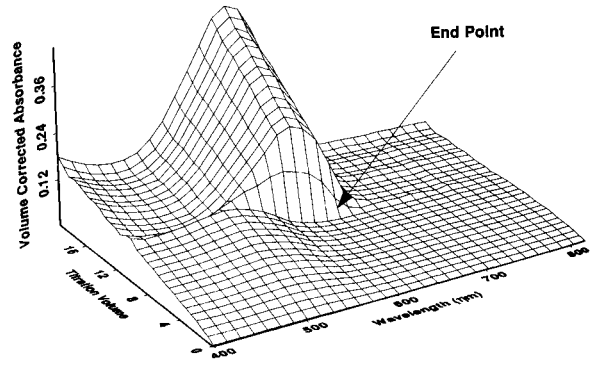


Fig. 4. Volume corrected absorbance spectra plotted as a function of titration volume.

Using a priori knowledge one could configure the number of units to use. Another method is to start with one unit in each layer and increment the units until no substantial decrease in error is observed. This method is similar to selecting the number of principal components. Figure 5 gives the root mean square prediction error as a function of the number of untis. The error can be measured with the calibration set or an external evaluation set. Minimizing the number of units is a useful heuristic.

A plot of RMS prediction error as a function of the number of epochs is given in Fig. 6 for both underdetermined and overdetermined data. Note that PCA training is less efficient in regard to the number of training epochs although it is

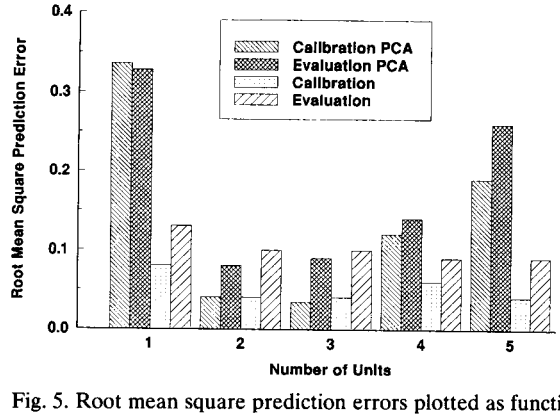


Fig. 5. Root mean square prediction errors plotted as function of the number of units for FLIN networks trained with PCA compressed and uncompressed data.

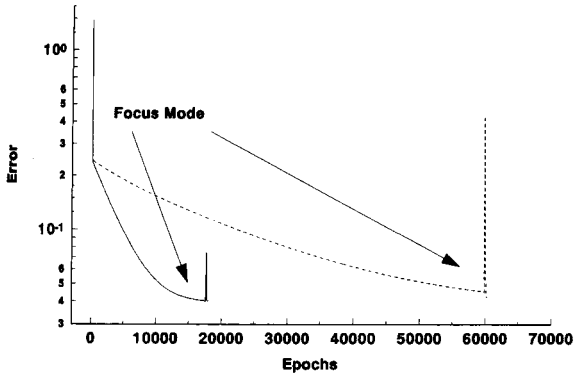


Fig. 6. Training error as a function of the number of training epochs for compressed and the full spectrophotometric data sets.

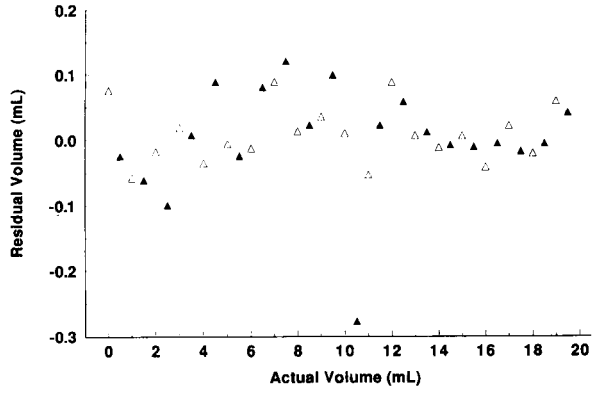


Fig. 8. Prediction residual plotted as function of actual volume for the full data set.

more efficient with respect to computational time. The peak in the figures indicates the transition to focus mode for which the overall training error increases as the units reconfigure themselves. This step involves units modeling different inputs.

The titration data is modeled by a two unit network. The number of K- and G-units in each layer is two. The number of input units is 211. The G-units function as output units, with only one G-unit firing at a time.

Figures 7A and 8A give the predicted titration volumes for both underdetermined and overdetermined data. The residuals are plotted in Figs. 7B and 8B. In both the predictions plots and the residual plots an outlier may be observed at 10.5 ml. This outlier was caused by lack of precision in the absorbance measurement and was verified to be incorrect by a second titration.

The advantage to local processing networks is information may be gained by examining the network. The loadings of the G-units and K-units may be directly observed. Note the K-vectors resemblance to the two litmus spectra in Fig. 9. These vectors furnish information regarding similarities within the data. The G-unit variable loadings in Fig. 10A and B are what would be expected from litmus absorbance spectra. There is not much difference between the PCA compressed and the uncompressed vectors.

A third study was run that decreased the threshold for adjustment of learning rates and training termination to a change in error of 1.0×10^{-5}. Training was completed after approxi-

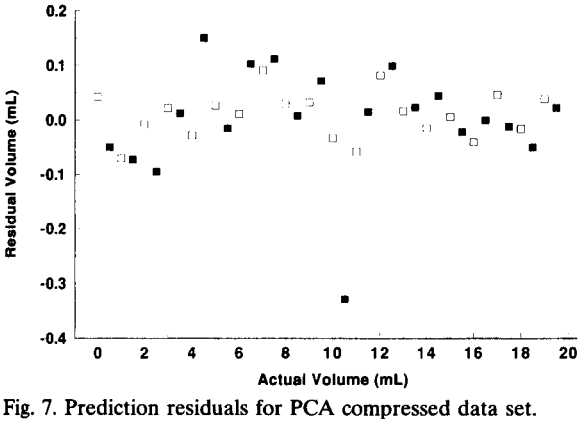


Fig. 7. Prediction residuals for PCA compressed data set.

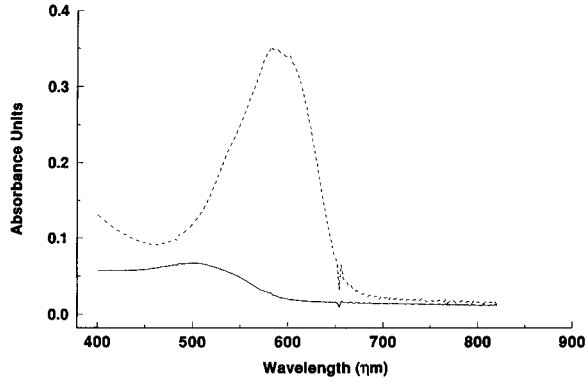


Fig. 9. Kohonen units for the uncompressed data set.

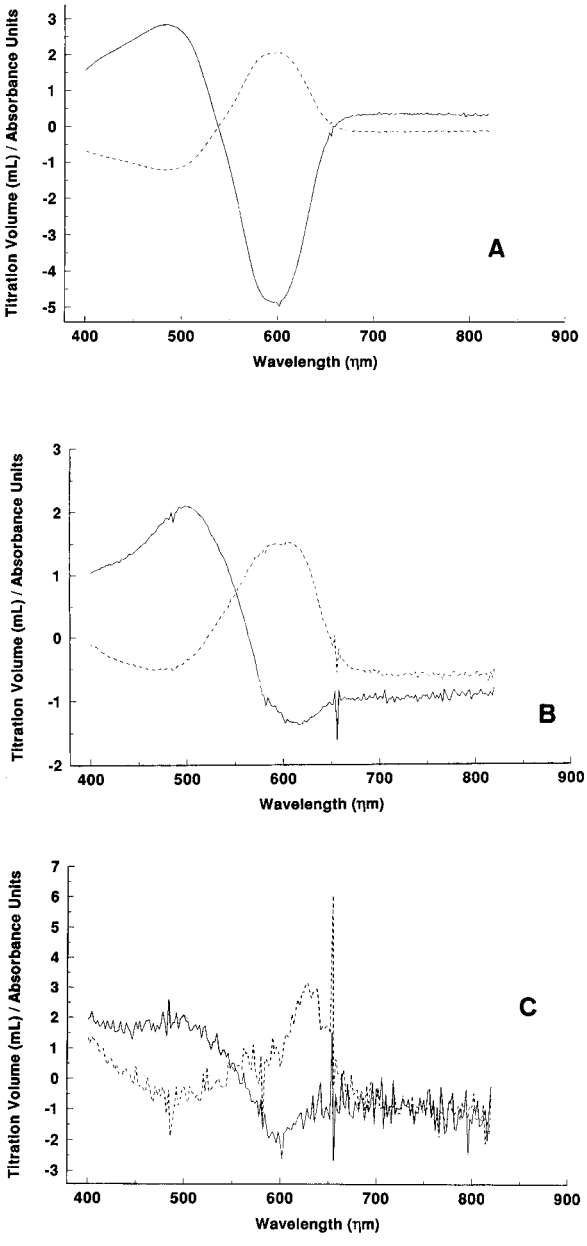


Fig. 10. (A) Grossberg units for the PCA compressed data. (B) Grossberg units for the uncompressed data, short training duration. (C) Grossberg units for the uncompressed data, long training duration.

mately 6.3×10^6 epochs. Note that overtraining causes small artifacts in the data to be of superfluous importance in Fig. 10C. With black box classifiers, overtraining may only be ascertained based on predictions. The RMS prediction error for this network was 0.031 for the calibration set and 0.097 for the evaluation set.

Conclusion

The FLIN network is a prototype network and many refinements are required before it can be used as a general purpose tool. The network uses a unique training method that enhances the relations between Kohonen and Grossberg layers. FLIN uses local processing which makes it a worthy research tool and probe of neural network learning. Non-linear modeling may open the door to many novel methods of analysis.

A future research goal is to develop an evolutionary algorithm for which K-units compete and optimize the number of units in the network during training. New constraints to prevent overfitting in delta learning methods may be studied with FLIN.

Hans Whittenburg is thanked for his help running experiments. Alan Hendricker, Peter Tandler and Busolo Wabuyele are thanked for their helpful comments and criticisms. Ohio University is thanked for their financial support through the Research Challenge program. James Y. Tong is thanked for the use of his instrument.

REFERENCES

1 P.B. Harrington, Chemom. Intell. Lab. Syst., 18 (1993) 157.
2 P.B. Harrington, J. Chemom., 5 (1991) 467.
3 P.J. Gemperline, J.R. Long and V.G. Gregoriou, Anal. Chem., 63 (1991) 2313.
4 K.L. Peterson, Anal. Chem., 64 (1992) 379.
5 R. Hecht-Nielsen, Appl. Opt., 26 (1987) 4979.
6 P.D. Wasserman, Neural Computing Theory and Practice, Van Nostrand Reinhold, New York, 1989, p. 61.
7 T. Kohonen, Self-organization and Associative Memory (Series in Information Sciences, Vol. 8), Springer Verlag, Berlin, 1988.
8 S. Grossberg, Progress in Theoretical Biology, Vol. 3, Academic Press, New York, pp. 51–141.
9 D.L. Massart and L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
10 D.E. Rumelhart, G.E. Hinton and R.J. Williams, in D.E. Rumelhart and J.L. McClelland (Eds.), Parallel Dis-

tributed Processing, MIT Press, Cambridge, MA, 1986, pp. 444–459.
11 T. Naes, T. Isaksson and B.R. Kowalski, Anal. Chem., 62 (1990) 664.
12 R. Danielsson and G. Malmquist, Chemom. Intell. Lab. Syst., 14 (1992) 115.
13 E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 2nd edn., 1991.
14 W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988, Chap. 10.

Analytica Chimica Acta, 277 (1993) 199–204
Elsevier Science Publishers B.V., Amsterdam

Some novel methods based on recursive optimal estimation. Applications to analytical chemistry

Banmei Yu

Department of Applied Physics, Changsha Institute of Technology, Changsha 410073 (China)

Menglong Li, Aiping Liu and Zhiliang Li

Department of Chemistry and Chemical Engineering, Hunan University, Changsha 410082 (China)

Leming Shi and Zhongxiao Pan

Institute of Chemical Metallurgy, Academia Sinica, Beijing 100080 (China)

(Received 3rd September 1992; revised manuscript received 6th January 1993)

Abstract

Two novel methods based on recursive optimal estimation were approached: fading Kalman filtering (FKF) and networked Kalman filtering (NFK). Fading Kalman filtering was employed to enhance overlapped spectroscopic resolution. Based on the nature of the Kalman filter, that the residual sequence is uncorrelated when the optimal gain is obtained, a new fading optimal adaptive algorithm is proposed and utilized. By on-line adjustment of the fading factor, the convergency and optimality of the Kalman filter were improved using measured outputs or estimated results, even in the presence of model errors and/or the effects of unmeasurable external disturbances. The FKF algorithm developed was applied to overlapped peak resolutions and gave good results. The networked Kalman filtering (NKF) of multiple models was found to have some advantages over the ordinary Kalman filtering and was used to detect concentrations of minor impurities of less than 1% of that of the major analyte.

Keywords: Optimization methods; Impurity detection; Kalman filtering; Multicomponent analysis; Overlapped peaks resolution; Recursive optimal estimation

The Kalman filter [1–3], an optimal least-squares filtering algorithm operating recursively on digital data, has found many applications in analytical chemistry (see some recent reviews [4–6]) including the deconvolution of overlapped responses in spectrometry [7–10], voltammetry [11], x-ray fluorescence [12], chromatography [13,14], etc. Some approaches employing Kalman filtering and/or factor analysis for treatment of data obtained from direct spectrofluorimetry [15–17] and from fluorescence detection in conjunction with chromatography (TLC, etc.) [18–21] were made to resolve overlapped peaks (curves) [15–17,19,20] background subtraction [18] and correction [21].

In recent years, several types of Kalman filters, classical or ordinary, adaptive, root-squares basic, and reduced or scalar filtering algorithms, were successfully used to resolve some problems occurring in analytical chemistry. In this paper, first, the fading Kalman filter (FKF) was studied and used for multi-component analysis; and second, networked Kalman filtering (NKF) was examined and used to detect small quantities of impurities.

Correspondence to: B. Yu, Department of Applied Physics, Changsha Institute of Technology, Changsha 410073 (China).

FADING KALMAN FILTERING – PRINCIPLE AND ALGORITHM

The Kalman filter is an optimal, recursive and digital filter which is based on a two-part model describing the system dynamics (state–space) and the measurement processing [2–6]. The two equations which describe the Kalman filtering model, and the classical Kalman filtering algorithm equations, are given in Table 1.

When the system model is accurate enough, CKF works well. However, if there is model noise error, CKF will not give acceptable results or a converged estimate. In order to prevent deconvergence, a fading Kalman filtering (FKF) algorithm is proposed by introducing a so-called fading or forgotten factor. That is to say, Eqn. 4 is replaced by Eqn. 10:

$$\mathbf{P}(k/k-1) = \mathbf{f}(k)\mathbf{F}(k, k-1)\mathbf{P}(k/k-1/k-1) \times \mathbf{F}(k, k-1) + \mathbf{Q}(k-1), \quad \mathbf{f}(k) \geqslant 1 \tag{10}$$

where $\mathbf{f}(k) \geqslant 1$ is defined as the fading factor, which pays more attention to the recent data than the past data. In this case, different weighting factors were given to new and old data. A greater weight was given to the new data and

TABLE 1

Classical Kalman filtering algorithm

System dynamic (state–space) model:

$$\mathbf{X}(k) = \mathbf{F}(k, k-1) * \mathbf{X}(k-1) + \mathbf{W}(k), \quad \mathbf{X}(k) \in \mathbb{R}^{(n)} \tag{1}$$

Measurement (process) model:

$$\mathbf{z}(k) = \mathbf{H}'(k) * \mathbf{X}(k) + v(k), \ \mathbf{Z}(k) \in \mathbb{R}^{(m)} \tag{2}$$

State estimation extrapolation (prediction):

$$\mathbf{X}(k/k-1) = \mathbf{F}(k, k-1)\mathbf{X}(k-1/k-1) \tag{3}$$

Covariance estimate extrapolation:

$$\mathbf{P}(k/k-1) = \mathbf{F}(k, k-1)\mathbf{P}(k-1/k-1)\mathbf{F}'(k, k-1) + \mathbf{Q}(k-1) \tag{4}$$

Innovation (scalar):

$$\mathbf{y}(k) = \mathbf{z}(k) - \mathbf{H}'(k)\mathbf{X}(k/k-1) \tag{5}$$

Kalman gain (vector):

$$\mathbf{G}(k) = \mathbf{P}(k/k-1)\mathbf{H}(k)[\mathbf{H}'(k)\mathbf{P}(k/k-1)\mathbf{H}(k) + \mathbf{R}(k-1)]^{-1} \tag{6}$$

State estimate update (filtering):

$$\mathbf{X}(k/k) = \mathbf{X}(k/k-1) + \mathbf{G}(k)\mathbf{y}(k) \tag{7}$$

Error covariance update:

$$\mathbf{P}(k/k) = [\mathbf{I} - \mathbf{G}(k)\mathbf{H}'(k)]\mathbf{P}(k/k-1) \tag{8}$$

or

$$\mathbf{P}(k/k) = [\mathbf{I} - \mathbf{G}(k)\mathbf{H}'(k)]\mathbf{P}(k/k-1)[\mathbf{I} - \mathbf{G}(k)\mathbf{H}'(k)] + \mathbf{G}(k)\mathbf{R}(k)\mathbf{G}'(k) \tag{9}$$

Definitions:

n	Number of estimated parameters or components
m	Total number of measurement points (e.g. wavelengths)
k	Number of the most recent measurement or indicating the independent variable
$\mathbf{X}(k)$	State vector ($n * 1$)
$\mathbf{F}(k,j)$	State transition matrix ($n * n$)
$\mathbf{W}(k)$	System (state) noise vector ($n * 1$)
$\mathbf{Q}(k)$	System (state) noise covariance matrix ($n * n$)
$\mathbf{z}(k)$	Measured responses (scalar, $1 * 1$, e.g. absorbances, fluorescences, etc.)
$\mathbf{H}'(k)$	Measurement function vector ($1 * n$)
$v(k)$	Measurement noise (scalar, $1 * 1$)
$\mathbf{R}(k)$	Measurement noise variance (scalar, $1 * 1$)
$\mathbf{P}(k/j)$	Prediction error covariance matrix ($n * n$) for $\mathbf{X}(k)$, based on measurements $\mathbf{z}(1), \ldots, \mathbf{z}(j)$
$\mathbf{G}(k)$	Kalman filtering gain weighting factor (vector, $n * 1$)
$\mathbf{I}$	Identity matrix ($n * n$)
0	Zero matrix ($n * n$) or zero vector ($n * 1$)
$'$	Transpose of matrix or vector

TABLE 2

Analytical results for five-dye mixtures [a]

Number	Amount taken (μg ml^{-1})					Amount found (μg ml^{-1})				
	TTZ	SSY	PNC	AMR	IDB	TTZ	SSY	PNC	AMR	IDB
52	3.0	8.0	1.0	9.0	2.0	3.157	7.942	1.026	8.810	2.017
53	4.0	6.0	2.0	8.0	4.0	4.023	5.936	1.943	7.815	4.122
54	5.0	5.0	3.0	6.0	5.0	5.118	5.176	3.107	6.002	5.157
55	6.0	4.0	3.0	5.0	6.0	6.192	4.155	2.876	5.026	5.863
56	7.0	3.0	4.0	4.0	7.0	6.850	2.973	4.058	4.237	6.822
57	8.0	2.0	5.0	2.0	8.0	7.782	2.161	4.833	2.129	8.057

[a] TTZ = Tartrazine; SSY = sunset yellow; PNC = carmine; AMR = amaranth; IDB = indigotin.

more of the information was extracted from the new data. So, the ability to trace data therefore was increased. The fading factor can be taken as either a constant or a variable. Obviously, if $\mathbf{f}(k)$ is always taken as one, $\mathbf{f}(k) = 1$, then FKF is reduced to CKF. The greater $\mathbf{f}(k)$ is, the bigger the weight is and, simultaneously, the larger the tracing noise.

In general, the fading factor should be properly selected by a compromise between the tracing ability and the tracing noise.

APPLICATIONS OF FKF – MULTICOMPONENT ANALYSIS

First, the FKF was applied to simulation experiments. When there are coefficients of error in the model and there are dynamic deviations in the system considered, a fading adaptive algorithm such as FKF with the optimal adaptive fading factor can eliminate the filtering residuals.

Second, the FKF was used to resolve overlapped spectra and for the simultaneous determination of multiple components in a synthetic mixture. By using FKF and spectrophotometry or spectrofluorometry, many multicomponent analyses were performed. For instance, some mixtures composed of two to six dyes were analyzed by UV–Vis spectroscopy with recoveries ranging from 96.5% to 108%. Some of the results for the determination of five dyes in a mixture are shown in Table 2. Table 3 shows some of the results for the resolutions of four amino acids in a mixture. The amounts found by FKF agree well with the amounts taken and those obtained by CKF, which are omitted here.

NETWORKED KALMAN FILTERING – PRINCIPLE AND ALGORITHM

Usually Kalman filtering employs only one set of unitary models composed of dynamics and measurement equations (Eqns. 1 and 2). In contrast, the parallel Kalman filter network [21,22] applies more than one set of models, in other words, the networked Kalman filtering (NKF) consists of several models of differing dimensionality having a similar formula as the equations mentioned above. Here, however, the vector, $\mathbf{H}'(k)$ or $\mathbf{X}(k)$, in each equation possesses various numbers of elements corresponding to various components based on each model. The number of components or the dimensionality of the various models, n_d, is given by 1, 2,..., etc. The corresponding measurement equation can be ex-

TABLE 3

Analytical results for four-amino acid samples

Number	Amount taken (μg ml^{-1})				Amount found (μg ml^{-1})			
	Tyr	Trp	Phe	Dhp [a]	Tyr	Trp	Phe	Dhp
142	10.0	8.0	10.0	20.0	10.52	7.94	10.38	19.56
143	15.0	7.0	20.0	15.0	14.70	7.07	19.46	15.24
144	20.0	6.0	20.0	15.0	20.26	6.03	22.26	14.79
145	10.0	5.0	30.0	10.0	10.63	4.85	28.57	10.65
146	15.0	4.0	30.0	10.0	15.38	4.12	29.35	10.47
147	20.0	2.0	40.0	8.0	19.87	1.93	41.35	8.26

[a] 3,4-Dihydroxyphenylalanine.

pressed as:

$$\mathbf{Z}_p(k) = \sum \mathbf{H}_i(k) * \mathbf{X}_i(k) + \mathbf{V}_p(k)$$
$$= \mathbf{H}'_p(k) * \mathbf{X}_p(k) + \mathbf{V}_p(k) \quad (11)$$

where $p = 1-n_d$ and $i = 1-p$. When the dimensionality, $n_d = 3$, for example, Eqn. 11 can be written as:

$$\mathbf{Z}_1(k) = \mathbf{H}_1(k)\mathbf{X}_1(k) + \mathbf{V}_1(k) \quad (12)$$

$$\mathbf{Z}_2(k) = \mathbf{H}_1(k)\mathbf{X}_1(k) + \mathbf{H}_2(k)\mathbf{X}_2(k) + \mathbf{V}_2(k) \quad (13)$$

$$\mathbf{Z}_3(k) = \mathbf{H}_1(k)\mathbf{X}_1(k) + \mathbf{H}_2(k)\mathbf{X}_2(k) + \mathbf{H}_3(k)\mathbf{X}_3(k) + \mathbf{V}_3(k) \quad (14)$$

For each model, the related innovation sequences, $\mathbf{y}_p(k)$ for the p-component model (in this case, $p = 1$–3) is calculated from Eqns. 11–14 and given by different values. Suppose three arbitrary but proper responses at the measurement points, designed $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$, are selected as the independent variables for the models, i.e. those hidden or fictitious variables composed of other unknown independent variables. A series of models of various dimensions, with the relevant forms of Eqns. 12–14, can be obtained for one, two and three independent variables, respectively. In NKF, these models are all employed to estimate the state parameters and to give the different innovation scalar quantities, $\mathbf{y}_1(k)$, $\mathbf{y}_2(k)$ and $\mathbf{y}_3(k)$. When only the pure compound is present as the primary component, then all the innovation series are white zero-mean gaussian noise series. Once an impurity is present as a secondary component, the first-order model will give a correlative sequence without zero mean, but the other innovation series $\mathbf{y}_2(k)$ and $\mathbf{y}_3(k)$ still behave as white, gaussian noise processes.

If the dimensionality or number of components, n_d, in the filtering models is equal to or greater than the number of analytes in the actual systems, the innovation sequences will behave as white and zero-mean gaussian noise; if n_d is less, they will not. The peak purity evaluation and the coexisting impurity detections are performed on the basis of the magnitude and changes in the innovation sequences. In practical research, the innovation sequences are not used directly, since there are many innovation series. The root mean squares (RMS) of the innovation sequence $\mathbf{y}(k)$ were, alternately, calculated by the following equation:

$$\mathrm{RMS}(\mathbf{y}(k)) = \mathrm{SQR}\left(\sum \mathbf{y}_p(k)/n_f\right) \quad (15)$$

or RMS ($\mathbf{y}_0(k)$) by using the orthogonal or perpendicular innovation, according to reference [22]. Where $n_f = n - p$ is the number of degrees of freedom of the models used or the number of measurement points used and the summation is made over all the points. Either RMS($\mathbf{y}(k)$) or RMS($\mathbf{y}_0(k)$) can be used to detect whether an impurity is present or not, i.e., to judge the peak is pure or not.

In NKF, since proper but arbitrary original responses are directly selected as the hidden (fictitious) components or the model functions, no hypothesis need be made on two-way data, such as spectral and chromatographic shapes.

APPLICATIONS OF NKF – SIMULATION RESULTS

For simplicity, it is assumed that a single gaussian-peak curve is an accepted approximation of a full complement of chromatographic peaks and a portion of spectroscopic curve. Severely overlapped classical and/or modified gaussian distribution curves were simulated as chromatographic domains and simultaneously strongly overlapped double gaussian curves were simulated for spectroscopic profiles. Then the resultant data of the two-way (chromatograms × spectra) matrix were produced and superimposed with normal-distribution random noise with different standard deviations (SD) of the sum of the maximum responses of all compounds.

In the simulation studies, the novel NKF is proved to be good for detecting impurity at fairly low levels, even at less than 1%. Some aspects and several examples will be given here.

In one simulation, the amounts of the minor component as impurity compound was 1% of the major compound; the chromatographic resolution was $R_{sc} = \Delta t_r/2(\sigma_{1c} + \sigma_{2c}) = 0.5$ and the spectral

resolution was $R_{sp} = \Delta\lambda_{max}/2(\sigma_{1sp} + \sigma_{2sp}) = 0.25$. A plot of RMS($y(k)$) against retention time is given in Fig. 1. It is clear that there are secondary or impurity compounds and these are located between points 30 and 50. In other words, the impurity of amounts as low as 1% of those of the major compound can be detected; and characterizing the location of the impurity is also realized by using the RMS innovation. In another study [22], only the results for a second component at levels as low as 1:10 or 10% were reported. Besides, based on principal component analysis (PCA) and factor analysis (FA), the minor compound(s) present in amounts of less than 5% do not induce a second significant eigenvector or that the detection limit for a second component will be in the range 2–10% of the total absorbance. For this reason it is not easy by FA to decide whether an eigenvalue is caused by small amounts, i.e., less than 2%, of an impurity or by the measurement noise based on some criterion. The moving fixed-size window (MFW)–evolutionary factor analysis (EFA) has been developed by Keller and Massart [26], which works very well in determining peak purity.

Fortunately, in this situation, the RMS sequences of innovation, can be useful for deciding whether an impurity is present or not and the "peak" of the RMS values of the innovation sequence can be employed to distinguish a signal produced by secondary impurities from that caused by noise and to give the location of impurity substance(s) in the system being observed.

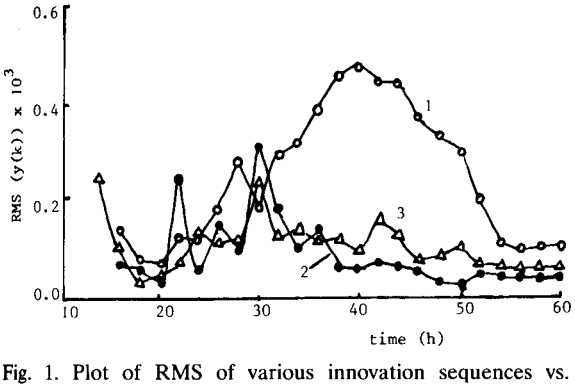


Fig. 1. Plot of RMS of various innovation sequences vs. retention time. Curves 1–3 stand for one-, two-, and three-dimensional (or -component) models, respectively.

From our primary results, we conclude that the NKF of multiple models has some advantages over the ordinary KF of a univariate model and over classical FA. NKF is able to evaluate peak purity and/or to detect whether a small impurity is present or not, under extremely high spectral similarity and quite low chromatographic resolution. Besides, the NKF technique can simultaneously perform data processing while more recent data is being acquired, therefore it may become an effective tool for on-line impurity detection and real-time quality control.

The authors gratefully acknowledge financial support from the National Natural Science and Open Laboratory Foundations as well as Hunan University and STAR Applied Chemistry Institute Research Foundations.

REFERENCES

1 R.E. Kalman, J. Basic Eng., 82 (1960) 35.
2 R.G. Brown, Introduction to Random Signal Analysis and Kalman Filtering, Wiley, New York, 1983.
3 A. Gelb (Ed.), Applied Optimal Estimation, M.I.T. Press, Cambridge, MA, 1974.
4 S.D. Brown, Anal. Chim. Acta, 181 (1986) 1.
5 S.C. Rutan, J. Chemom., 1 (1987) 7.
6 Z.L. Li and L.M. Shi, ICCCRE, 8th, Beijing (1987) D-16; Z.L. Li, L.M. Shi, M.L. Li and R.Q. Yu, Proc. Nat. Conf. Comput. Anal. Chem.: Collection of Plenary Lectures Beijing, Chin. Chem. Soc., 2nd (Lushan, 1988), No. 5, pp. 1–18.
7 H.N.J. Poulisse, Anal. Chim. Acta, 112 (1979) 361.
8 A. van Loosbroek, H.J.G. Debets, and P.M.J. Coenegracht, Anal. Lett., 17 (1984) 779.
9 Z.L. Li and L.M. Shi, J. Environ. Sci. (Beijing), 9 (1988) 56.
10 Z.L. Li, L.M. Shi, M.L. Li, H.G. Lin and R.Q. Yu, Acta Chim. Sinica, 48 (1990) 1101.
11 T.F. Brown and S.D. Brown, Anal. Chem., 53 (1981) 1410.
12 Z.L. Li, L.M. Shi, J. Xu, M.S. Zhang, Q.G. Wang and R.Q. Yu, Anal. Chim. Acta, 248 (1991) 257.
13 Y. Hayashi, T. Shibazaki, R. Matsuda and M. Uchiyam, J. Chromatogr., 407 (1987) 59.
14 D.D. Gerow and S.C. Rutan, Anal. Chim. Acta, 184 (1986) 53.
15 S.C. Rutan, D.D. Gerow and G. Hartmann, Chemom. Intell. Lab. Syst., 3 (1988) 61.

16 Z.L. Li and R.Q. Yu, Acta Chim. Sinica, 48 (1990) 1018.
17 Z.L. Li, L.M. Shi, M.L. Li and R.Q. Yu, Chem. J. Chin. Univ., 11 (1990) 245.
18 D.D. Gerow and S.C. Rutan, Anal. Chim. Acta, 184 (1986) 53.
19 S.C. Rutan, Chemom. Intell. Lab. Syst., 6 (1989) 191.
20 S.D. Brown, Chemom. Intell. Lab. Syst., 10 (1990) 87.
21 P.D. Wentzell and S.J. Vanslyke, Anal. Chim. Acta, (1992) in press.
22 S.J. Vanslyke and P.D. Wentzell, Anal. Chem., 63 (1991) 2512.
23 E.R. Malinnowski and D.G. Howery, Factor Analysis in Chemistry, Wiley, New York, 1980.
24 M.A. Sharaf and B.R. Kowalski, Anal. Chem., 53 (1981) 518; P.J. Gemperline, J. Chem. Inf. Comput. Sci., 24 (1984) 206; Anal. Chem., 58 (1986) 2656.
25 B.G.M. Vandeginste, R. Essers, T. Bosman, J. Reijnen and G. Kateman, Anal. Chem., 57 (1985) 971; B.G.M. Vandeginste, W. Deerks and G. Kateman, Anal. Chim. Acta, 173 (1985) 253; W. Lindberg, J. Ohman and S. Wold, Anal. Chem., 58 (1986) 299.
26 H.R. Keller and D.L. Massart, Anal. Chim. Acta, 246 (1991) 379; M. Maeder, Anal. Chem., 59 (1987) 527.

Analytica Chimica Acta, 277 (1993) 205–214
Elsevier Science Publishers B.V., Amsterdam

Multivariate decision and detection limits

Anita Singh

Lockheed Environmental Systems & Technologies Co., Environmental Sciences & Technologies Division, 980 Kelly Johnson Drive, Las Vegas, NV 89119 (USA)

(Received 3rd September 1992)

Abstract

Principal component analysis (PCA) is used to develop an approach for estimating multivariate decision and detection limits (MDDLs) for gas chromatography–mass spectrometry studies where the instrument response is multivariate in nature. Many definitions and estimators have been published for univariate responses. In this article we extend these ideas to the multivariate case. When the first principal component explains most of the variation contained in the data, it may be used to express the multicomponent instrumental response as a univariate composite signal representing all of the monitored ions associated with the analyte of interest. The first principal component of these ions has been used to derive decision and detection limits through two approaches. Back-transformation of this composite instrument response to the original response variables will yield multivariate decision and detection limits for all of the ions considered simultaneously.

Keywords: Mass spectrometry; Principal component analysis; Calibration model; Detection limits

In environmental chemical measurement it is necessary to detect low-level concentrations of hazardous compounds in various media. The instrument response resulting from the analysis of samples or standards that have low concentrations may be difficult to distinguish from the inevitable background noise. Uncontrolled noise factors operate whether or not the analyte of interest is present. These uncertainties are addressed in this paper for multicomponent instrument responses that are characteristic of gas chromatography–mass spectrometry (GC–MS). The decision and detection limits (DDLs) are often expressed in terms of the analyte concentrations. In practice, however, a chemist observes the instrumental response first. Therefore, these limits are determined in the signal space first and then transformed to the concentration domain by means of a suitable calibration model representing the relationship between the instrumental response and the analyte concentration. Thus, the problem is twofold: (1) obtain an appropriate calibration model for established standards using suitably designed experiments, and (2) develop appropriate procedures to estimate the DDLs using this calibration curve.

Assessing the DDLs of an instrument is not an easy task. There is much controversy in the chemical literature regarding the definition of these limits. Because these limits indicate the performance of an instrument at detecting low-level concentrations, an analytical chemist should be clear about the method used in the derivation of these DDLs. In this paper, we will not try to distinguish between the instrument detection limits (based on standards) and the method detection limits (based on spiked samples for a particular matrix). It is assumed that the chemist using these procedures will differentiate between these limits accordingly.

Correspondence to: A. Singh, Lockheed Engineering and Sciences Co., Environmental Programs Office, 980 Kelly Johnson Drive, Las Vegas, NV 89119 (USA).

Currie [1] used the univariate instrumental response corresponding to blank analyses (in which the substance sought is absent) to establish univariate DDLs. He defined these limits in terms of the instrumental responses only. To define these limits in terms of analyte concentration of interest, it is necessary to rely on a calibration curve relating the instrument responses to concentration levels in established standards.

Hubaux and Vos [2] used least squares regression of the univariate response onto the analyte concentration of a certain number of calibration standards. They used this regression function as an estimate of the true calibration curve. It is anticipated that any new standard will yield a signal falling in the neighborhood (regression band) of this regression line. Using this calibration curve, they defined the decision limit in the response domain and the detection limit in the concentration domain. Also Garner and Robertson [3] reviewed various methods of estimating these DDLs for the univariate case.

In this article, we extend these ideas to the multivariate case, where the instrumental response is inherently multivariate such as in GC–MS studies. Principal component analysis (PCA) [4] plays an increasingly important role in recent developments in chemometric literature. Due to the complex nature of the multicomponent analytical data obtained in GC–MS studies, in practice only one or two most abundant ions are used to quantitate the analyte concentration of interest. Information on the other ions is usually applied only for identification and confirmation. Due to the high precision in the measurement process in calibration experiments of the GC–MS studies, the first principal component of the relevant monitored ions typically accounts for most of the variation ($\geqslant 99\%$) in the data. Without loss of significant amount of information, the first principal component can be used to represent the composite instrument response, which in turn can be used to quantitate the analyte of interest. Moreover, in calibration experiments, the magnitude of the last few principal components can be used to determine how well the first few principal components fit the observations. Using these first few principal components, one can back-transform the principal component scores to the original response variables (if desired). Delaney [5] used PC to obtain these limits using the Hubaux and Vos approach.

Least squares regression has been used to obtain a calibration function for the composite instrument response onto the analyte concentration in established standards. Quantitation of the analyte concentration based on this calibration curve uses information on all of the relevant major ions and, in general, yields more precise estimates of the analyte concentration than the classical univariate techniques. Moreover, this calibration curve can be used to define the DDLs in terms of the analyte concentration. To obtain a calibration curve that can be used efficiently to estimate these limits, it is important to follow an appropriate design to carry out these experiments. Enough samples should be included in the design of experiments, so that one can obtain reasonably precise estimates of the unknown response parameters at low concentration levels. Moreover, because the principal components are linear functions of the original response variables, an appeal to the central limit theorem helps to justify the normality assumption for the principal components, even when the original response variables are not Gaussian. This justifies adaptations of the Currie and the Hubaux–Vos type DDLs, both of which assume Gaussian distribution for the instrumental response at all levels of the analyte concentration. In this article, we derive these limits by using two approaches, those of Currie and of Hubaux and Vos. Comparisons have been made between the results obtained by using (1) only the one or two most abundant ions (univariate approach) and (2) the composite response based on the first principal component (multivariate approach). The scores for the first principal component are transformed back to the original responses to give rise to MDDLs. It is important to transform the composite DDLs to MDDLs to ensure that these MDDLs lie within the predefined identification region for the target analyte to confirm its presence. Calibration data of 2,3,7,8-TCDD obtained according to the USEPA contract IFB WA 84-A002 have been used in this study. The pre- and post-transformed

data are in close agreement and are summarized in Table 3.

MATHEMATICAL FORMULATIONS

Let $\boldsymbol{R}' = (r_1, r_2, \ldots, r_p)$ be the multicomponent response representing the relative peak areas of $p(\geq 1)$ ions of interest included in a particular GC–MS study. Let $x_1, x_2, \ldots, x_n$ be the n concentration levels in established standards included in the study. The observed response matrix thus obtained can be expressed as follows:

$$\begin{bmatrix} r_{11}, & r_{12}, & \ldots, & r_{1p} \\ r_{21}, & r_{22}, & \ldots, & r_{2p} \\ & \cdot & & \\ & \cdot & & \\ r_{n1}, & r_{n2}, & \ldots, & r_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_1' \\ \boldsymbol{R}_2' \\ \cdot \\ \cdot \\ \boldsymbol{R}_n' \end{bmatrix} \quad (1)$$

The dispersion matrix for the p-responses is given by

$$\mathbf{S} = \sum_{i=1}^{n} (\boldsymbol{R}_i - \bar{\boldsymbol{R}})(\boldsymbol{R}_i - \bar{\boldsymbol{R}})' / (n-1)$$

where

$$\bar{\boldsymbol{R}} = \sum_{i=1}^{n} \boldsymbol{R}_i / n$$

Principal component analysis

Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the p-eigenvalues of the dispersion matrix $\mathbf{S}$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ be the corresponding eigenvectors. The p-principal components are given by $y_i = \mathbf{e}_i' \boldsymbol{R}$; $i = 1, 2, \ldots, p$.

The proportion of variation explained by the first k principal components is $A\nu_k = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$; $k = 1, 2, \ldots, p$. In practical applications, when $A\nu_1$ is large enough (≥ 0.95), without loss of a significant amount of information contained in the data, one can use the first principal component y_1 as the composite instrument response representing all of the p ions included in the study. Estimates of the DDLs can be obtained in terms of this composite response, which in turn can be back-transformed to the p-dimensional original response vector $\boldsymbol{L}$ giving rise to the MDDLs. If these MDDL vectors fall in the p-dimensional identification region $\boldsymbol{I}$ and meet all of the identification criteria set by the researchers, then the presence of the analyte in a particular field-sample can be confirmed. The back-transformation can be performed by expressing the response vector $\boldsymbol{R}_j$ as a linear combination of the PCs as follows

$$\boldsymbol{R}_j = \sum_{i=1}^{k} y_{ij} \mathbf{e}_i;\ j = 1, 2, \ldots, n, \text{ and } k = 1, 2, \ldots, p \quad (2)$$

The magnitudes of the last few principal components determine how well the first few principal components fit the observations. That is $\sum_{i=1}^{k} y_{ij} \mathbf{e}_i$ differs from $\boldsymbol{R}_j$ by $\sum_{i=k+1}^{p} y_{ij} \mathbf{e}_i$, when the first k principal components are used to fit the observations. In the dioxin study considered in this article, the first principal component alone was adequate to reproduce the original response matrix.

Calibration model

A calibration curve is routinely used for quantitation of an analyte of interest. Moreover, the estimates of the DDLs are first obtained in the response space and converted to the concentration domain through an appropriate calibration model. Thus, it is important to obtain a calibration model that is suitable for both purposes. The general calibration problem can be described by the following model

$$y(x) = \beta_0 + h(x, \boldsymbol{\beta}) + \epsilon \quad (3)$$

where $y(x)$ is the composite observed response at concentration level x, β_0 is the expected background noise, $\boldsymbol{\beta}$ is the vector of regression parameters, and ϵ represents the random measurement error. Routinely, these measurement errors are assumed to be independently and identically distributed as $N(0, \sigma^2)$ at all concentration levels. If the normality assumption is violated, a suitable Box-Cox type transformation given by Johnson and Wichern [6] can be used to transform the data to near-normality. Another assumption which is often violated is the homogeneity of variances. Variances are about the same at low concentrations, but are statistically significantly different at higher concentrations. Statistical tests

should be performed to test the homogeneity of variances that in turn require sufficient number of observations at each concentration to be included in the calibration experiment. An estimate of the functional form $h(x, \beta)$ of the model is obtained using the ordinary least squares (OLS) regression for homoscedastic responses or the weighted least squares (WLS) regression for heteroscedastic responses as described in Draper and Smith [7]. In order to get the best possible calibration model, it is important to design the experiments with clear objectives about its usage in mind. If this model has to be used to estimate the DDLs, then enough experiments at low concentrations should be included in the study. If this curve is also to be used for quantitation and prediction purposes, then the researcher should include enough experiments covering the practical range in the concentration domain. With experiments that are appropriately designed, more meaningful and precise results can be extracted from the available data.

The simplest calibration model is given by

$$y(x) = \beta_0 + \beta_1 x + \epsilon \tag{4}$$

where the estimates of β_0 and β_1 are obtained by using the OLS or the WLS regression on n calibration experiments. The estimates using the OLS regression are given by

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-2)$$

with

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i; \quad i = 1, 2, \ldots, n \tag{5}$$

Chemists measure the instrumental response first and then estimate the analyte concentration through a suitable calibration model. Thus, the inverse calibration problem is to estimate the analyte concentration x_0 for an observed response y_0 (based upon some $r \geq 1$ replicates). The classical point estimator $\hat{x}_0$ of x_0, obtained by Eqn. 4 is also a maximum likelihood estimator (MLE) of x_0. This estimator, however, has infinite mean square error (see Shukla [8]). Because no other uniformly better (see Krutchkoff [9] and Wiliams [10]) estimator is available in the literature, this classical MLE $\hat{x}_0$ is widely used in practice. Moreover, the mean $E[\hat{x}_0]$ and variance $V[\hat{x}_0]$ of $\hat{x}_0$ are infinite. Therefore, the interval estimate (x_L, x_U) for x_0 with confidence coefficient $(1-\alpha)$ is obtained by using the regression band around the model given by Eqn. 4. This $(1-\alpha)100\%$ confidence band is represented by the two curves given by Eqn. 6, for various values of the analyte concentration x within the region used in the calibration experiment.

$$\left.\begin{matrix} y_U(x) \\ y_L(x) \end{matrix}\right] = \hat{y}(x) \pm \hat{\sigma} t \left[(x - \bar{x})^2 / s_{xx} + 1/r + 1/n \right]^{1/2} \tag{6}$$

where $\hat{y}(x)$ is the average composite response based upon r replicates at concentration x.

The corresponding confidence interval for a particular analyte concentration x_0 within the experimental region is given by (x_L, x_U), where

$$\left.\begin{matrix} x_U \\ x_L \end{matrix}\right] = \hat{x}_0 + \left[(\hat{x}_0 - \bar{x}) g \pm \left(t\hat{\sigma} / \hat{\beta}_1 \right) \left[(\hat{x}_0 - \bar{x})^2 / s_{xx} + (1-g)(1/n + 1/r) \right]^{1/2} \right] / (1-g) \tag{7}$$

with $g = t^2 \hat{\sigma}^2 / (\hat{\beta}_1 s_{xx})$, $s_{xx} = \sum (x_i - \bar{x})^2$, and t is the student's t-value with $(n-2)$ degrees of freedom, and is chosen such that $P[y_L < y_0 < y_U] = 1 - \alpha$, for a chosen level of significance α. The graphical representation of these expressions is given in Fig. 1.

We now define some statistical terms. Suppose that a chemist has analyzed a sample and observed the (composite) response y which is assumed to be normally distributed with probability density function given by

$$f_x(y) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-(y - \mu_x)^2 / 2\sigma_x^2 \right] \tag{8}$$

where x represents the amount of the analyte

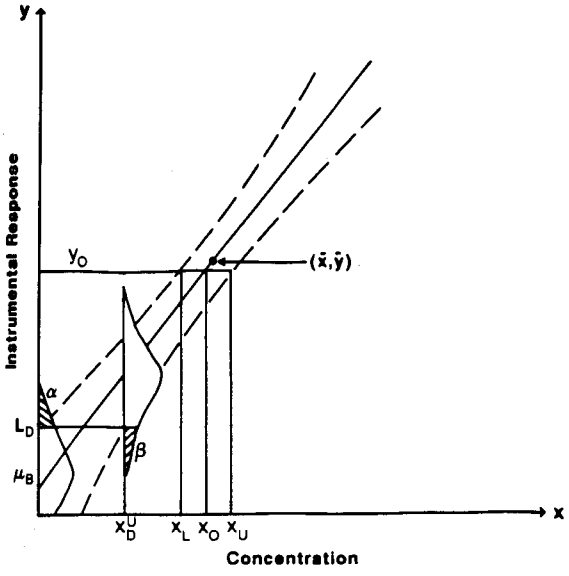


Fig. 1. Definitions of (a) the limit of decision L_D; and (b) the limit of detection x_D^U.

present in the sample, μ_x and σ_x^2 represent the corresponding mean and variance of the instrument response y at concentration level of x. Unless otherwise necessary, we do not use the subscript x in the following discussions. In practice, the random background noise present during the analytical process operates, whether or not the analyte is present.

This noise gives rise to the instrument response y_B distributed as $N(\mu_B, \sigma_B^2)$, even when the analyte sought is absent from the sample. The parameters μ_B and σ_B^2 are estimated from several replicates of blank samples. The central problem that the analyst now faces is to decide whether the instrumental response y is a result of the presence of this analyte or simply represents the background noise (or an analytical blank). In statistical terminology, an analyst wants to choose between the following two hypotheses:

H_0: the analyte sought is absent, i.e., $\mu = \mu_B$

H_1: the analyte sought is present, i.e., $\mu > \mu_B$

While resolving these uncertainties, the analyst is likely to commit the following two types of errors in detection with probabilities α and β respectively.

Type 1 error. Reject H_0 when H_0 is true, i.e, identify the background signal as the sample signal; this error constitutes a false positive.

Type 2 error. Reject H_1 when H_1 is true, i.e., identify the sample signal as the background signal; this error constitutes a false negative.

$$\alpha = P(\text{Type 1 error}) = P(y > y_2 \mid H_0)$$

$$= \int_{y_2}^{\infty} f(y_B)\, dy_B \qquad (9)$$

$$\beta = P(\text{Type 2 error}) = P(y < y_1 \mid H_1)$$

$$= \int_{-\infty}^{y_1} f(y_c)\, dy_c$$

where y_c is the instrument response with mean μ_c and variance σ_c^2 corresponding to an amount c of the analyte found in the sample. As the amount c decreases, it becomes more difficult to distinguish between the true response y_c and the background response y_B. The detection problem that the chemist now faces is how to resolve the differences between y_B and y_c while keeping the error probability α and β acceptably small. Throughout the rest of the paper, we will be making the assumption that the response standard deviations at low concentrations are equal, i.e., $\sigma_B \approx \sigma_{L_c} \approx \sigma_{L_d}$. These ideas are summarized in Fig. 2.

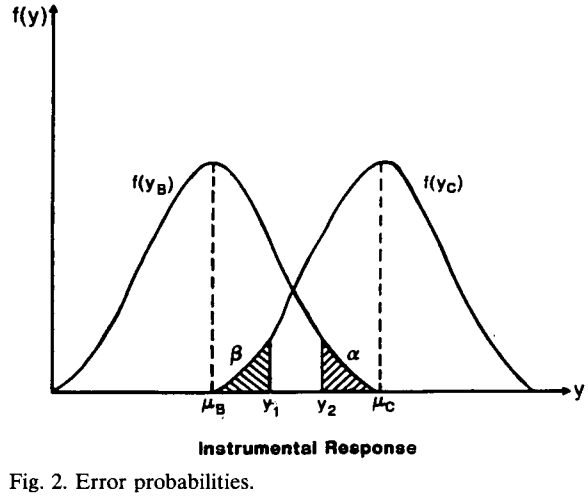


Fig. 2. Error probabilities.

Currie approach

Assuming normality of the instrument responses, Currie [1] defined $y_2 = \mu_B + z_\alpha \sigma_B$ as the decision limit L_c for $\alpha = 0.0013$ and $z_\alpha = 3.0$. Statistically, L_c represents the lowest meaningful (composite) response significantly different from the null response. Kaiser [11], defined L_c as the limit of detection. Also International Union for Pure and Applied Chemistry (IUPAC) [12] has recommended L_c as the limit of detection. However, for samples with the instrumental responses centered around L_c, the Type 1 error probability α is small (≈ 0.0013), i.e., signals larger than L_c represent the presence of the analyte with great precision, but the Type 2 error probability β is at least 0.50, i.e., signals smaller than L_c can be interpreted as the absence (false negative) of the compound with poor precision, which is unacceptable. Thus L_c cannot be used as a criterion of analyte detection for an analytical procedure. The response $L_c + z_\alpha \sigma_B$ becomes Currie's detection limit L_d for $\alpha \approx 0.0013$. Let $y_1 = L_d - z_\beta \sigma_B$. For instrumental responses centered around L_d, a response smaller than y_1 will be treated as the null response with β as the probability of false negatives. For $\alpha \approx \beta \approx 0.0013$, both y_1 and y_2 coincide at L_c. These limits are expressed in Fig. 3. Thus a signal y greater than L_d gets detected with high probability of at least 0.9987, whereas signals smaller than L_d also get detected with lesser accuracy. For example, a signal of magnitude $L_c + 1.96\sigma_B$ gets detected with probability 0.975. Thus far, L_c and L_d have been defined in terms of the composite instrument responses. In practice, the researcher may be interested in obtaining these limits in terms of the original p-responses, which are obtained by back-transforming the principal components to the original p-responses. These limiting response vectors are given as follows:

$$L'_c = \sum_{j=1}^{k} L_{cj} e_j = (L_{1c}, L_{2c}, \ldots, L_{pc})$$
$$L'_d = \sum_{j=1}^{k} L_{dj} e_j = (L_{1d}, L_{2d}, \ldots, L_{pd}) \quad (10)$$

where k is the number of PCs used to reproduce the observed response matrix given by Eqn. 1. The response vectors L_c and L_d should lie within the predefined p-dimensional identification region I to confirm the presence of the particular analyte of interest. These limits are defined in the response space, and by means of the calibration model, one can obtain the corresponding estimates of the DDLs in the concentration domain which we denote by x_c and x_d respectively. In practice x_c and x_d are estimated in more than one way. Different methods often lead to significantly different estimates of these limits. One commonly used approach (see Miller and Miller [13]) is to use the following equations to obtain the point estimates of x_c and x_d:

$$x_c^P = (L_c - \hat{\beta}_0)/\hat{\beta}_1$$

and

$$x_d^P = (L_d - \hat{\beta}_0)/\hat{\beta}_1$$

However, statistically the lowest analyte content that could be distinguished from being the null content is x_c^U (and not x_c^P), where x_c^U and x_d^U are the corresponding upper fiducial limits given by Eqn. 7 for x_c and x_d respectively. We recommend the use of x_c^U and x_d^U as the estimates of DDLs in the concentration space.

Currie [1] used normal distribution in the derivation of these DDLs, assuming that μ_B and σ_B^2 are known in advance. However, in practice

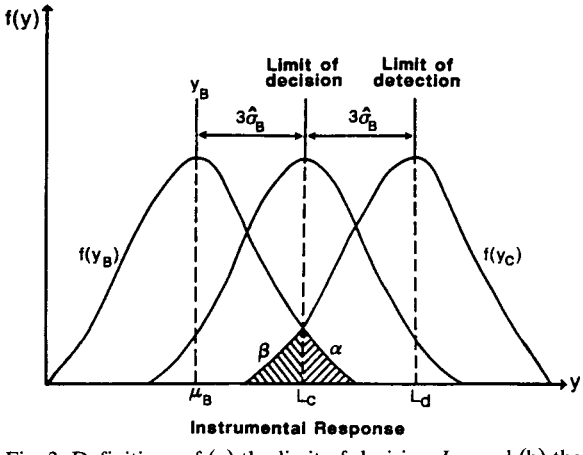


Fig. 3. Definitions of (a) the limit of decision L_c; and (b) the limit of detection L_d.

these parameters have to be estimated from some b number of blank samples. Therefore, it is more appropriate to use the student's t-distribution with $(b-1)$ degrees of freedom, rather than the conventional normal distribution in the derivation of these limits. Using student's t-distribution, Currie-type DDLs in the response domain are given by the following equations:

$$L_c = \hat{\mu}_B + t_{b-1}\hat{\sigma}_B\sqrt{1+1/b}$$

and

$$L_d = L_c + t_{b-1}\hat{\sigma}_B\sqrt{1+1/b}$$

where t_{b-1} is the tabulated t-value with $(b-1)$ degrees of freedom for some significance level of α.

Hubaux-Vos approach

Hubaux and Vos [2] used the regression band around the calibration curve given by Eqn. 4 to derive the DDLs. They defined the decision limit in the response space and the detection limit in the concentration domain. The composite decision limit L_D can be obtained from Eqn. 6 by letting $r \to \infty$ in equation $L_D = y_U(0)$, with $\hat{y}(0) = \hat{\beta}_0$. The corresponding estimate of the detection limit x_D is given by x_D^U of Eqn. 7 with $\hat{x}_0 = (L_D - \hat{\beta}_0)/\hat{\beta}_1$. Thus, for appropriate values of α and β, we have $P(y_B > L_D \mid \mu \approx \hat{\mu}_B) = \alpha$ and $P[y_c < L_D \mid \mu \approx \hat{y}(x_D^U)] = \beta$ (Fig. 1). An instrumental response greater than L_D represents a response statistically significantly different from the null response with a significance level of α. At concentration c, we have $P(y_c > L_D \mid H_1) \geq 1 - \beta$. The decision limit vector L_D in the original response space can be obtained by using Eqn. 10.

In most applications, the instrumental responses are heteroscedastic, therefore, the WLS regression [7] should be used rather than the OLS regression to obtain the calibration model represented in Eqn. 4. But in practice, since enough experiments are not available at each of the concentration levels included in the calibration experiments, one is forced to use the OLS regression to obtain Eqn. 4. This overestimates the error variance σ^2, which in turn yields inflated values of the DDLs when using the Hubaux and Vos approach.

EXPERIMENTAL

A GC–MS instrument was calibrated for 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) through three injections each of five standards. This experiment involved one instrument at a single laboratory. The calibration was performed as described in USEPA contract IFB WA 84-A002 for determining 2,3,7,8-TCDD in soils and sediments. The five standard solutions were provided by the USEPA. Each of these solutions contained the internal standards, $^{13}C_{12}$-labeled 2,3,7,8-TCDD, at a concentration of 1 ng μl^{-1}. These solutions contained unlabeled 2,3,7,8-TCDD at concentrations of 0.2, 1.0, 5.0, 20.0, and 40.0 ng μl^{-1}, respectively. These concentrations are equivalent to 50-μl concentrated extracts of 10-g samples containing 1, 5, 25, 100, and 200 ppb, respectively. Through selected ion monitoring, responses were observed for the mass-to-charge (m/z) ratios 257, 320, 322, 328, 332, and 334. The 2,3,7,8-TCDD yielded responses at m/z 320 and 322, with a fragment at m/z 257, while the internal standard is observed at m/z 332 and 334. A surrogate compound, 2,3,7,8-TCDD labeled with ^{37}Cl was present in the calibration also, with response at m/z 328, but surrogate information is used in quality control for sample analysis and is not part of our approach for estimating instrument detection limits.

The least squares regression model (Eqn. 4), gives the most precise results when the measured instrumental responses correspond to a point close to the centroid of this regression line. Because the main objective of this study is to estimate MDDLs, only solutions with low concentration levels of 0.2, 1.0, and 5.0 ng μl^{-1} of unlabeled 2,3,7,8-TCDD are used to develop the calibration model (Eqn. 4). The model thus obtained might not be ideal for general purpose calibration over the whole practical range of the analyte concentration.

Six ($b = 6$) available reagent blanks are also included in the study. There are six observations at a concentration level of 0.2 ng μl^{-1}, three from the initial calibration and three from the continuing calibration. In order to avoid the effect of the instrumental settings, the statistical

analysis is carried out on the relative instrumental responses, rather than on the original mass counts. The relative responses for m/z 257, 320, and 322 are given as follows: r_1 = (peak area at m/z 257)/D, r_2 = (peak area at m/z 320)/D, r_3 = (peak area at m/z 322)/D, where D = (peak area at m/z 322 + peak area at m/z 334).

TABLE 2

Hubaux-Vos approach using 98% confidence band

	PC analysis	322 alone	320+322
L_D (decision limit)	0.03966	0.03371	0.05295
x_D^U (detection limit) (ng μl^{-1})	0.1157	0.1156	0.1115
x_D^P (point estimate) (ng μl^{-1})	0.0586	0.0586	0.0565

RESULTS AND DISCUSSION

The DDLs are obtained using the conventional univariate as well as the proposed multivariate approaches. The univariate DDLs have been obtained in two ways (1) using the most abundant ion 322 and (2) using the sum of the two most abundant ions 320 and 322. PC analysis has been carried out on relative response vectors $\boldsymbol{R}_i$; $i = 1, 2, \ldots, 18$. The three eigenvalues of the variance covariance matrix **S** are 1.51077, 0.00015, and 0.0000285. The first PC y_1 alone explains 99.987% of the variation contained in the data. The calibration curve using this PC is given by $y_1 = 0.0005761 + 0.6662x$ with the correlation coefficient $r = 0.9991$. The calibration curve using response y at m/z 322 is given by $y = 0.004643 + 0.4959x$ with $r = 0.9991$; whereas the model based on the sum y of the responses at m/z 320 and 322 is given by $y = 0.003078 + 0.8824x$ with $r = 0.9991$. Each of the models fits the data equally well. The DDLs using these models are summarized in Table 1 (Currie approach) and Table 2 (Hubaux and Vos approach). The units used for all of the estimates in the concentration domain are in ng μl^{-1}. Regression bands with a confidence coefficient of 0.98 and a t-value of 2.5835 (with 16 degrees of freedom) have been obtained for each of the three models. These confidence bands have been used in estimating the DDLs in the concentration domain for both of the approaches.

In this study, from Tables 1 and 2, it is obvious that both of the multivariate as well as the univariate approaches produced similar results. However, one of the advantages of using the multivariate PCA approach is that, by means of back-transformation, the composite instrument response can be transformed to the original p monitored ions (here $p = 3$). The pre- and post-transformed data based upon the first principal component is given in Table 3. Using back-transformation, the following multivariate DDLs are obtained: $\boldsymbol{L}'_c$ = (0.00628, 0.011, 0.01412); $\boldsymbol{L}'_d$ = (0.00995, 0.01744, 0.02239) original Currie ap-

TABLE 1

Currie approach [a]

	Original approach using normal distribution			Modified approach using Student's t-distribution		
	PCA	322 alone	320 + 322	PCA	322 alone	320 + 322
$\hat{\mu}_B$	0.00787	0.00609	0.00954			
$\hat{\sigma}_B$	0.00370	0.00316	0.00447			
L_c	0.01897	0.01558	0.02296	0.02132	0.01759	0.02579
x_c^P (ng μl^{-1})	0.0276	0.0220	0.0226	0.0312	0.0260	0.0258
x_c^U (ng μl^{-1})	0.0850	0.0795	0.0780	0.0886	0.0835	0.0812
L_d	0.03007	0.02507	0.03637	0.03477	0.02908	0.04205
x_d^P (ng μl^{-1})	0.0442	0.0412	0.0377	0.0514	0.0492	0.0442
x_d^U (ng μl^{-1})	0.1016	0.0984	0.0930	0.1085	0.1064	0.0994

[a] The t-value used in the blank estimates of the modified Currie approach is $t_5 = 3.3649$.

TABLE 3

Pre- and post-transformed relative responses

Concentration (ng μl^{-1})	Observed ion, 257 r_1	Back-trans. ion, 257 r_1	Observed ion, 320 r_2	Back-trans. ion, 320 r_2	Observed ion, 322 r_3	Back-trans. ion, 322 r_3
0.0	0.006006	0.003440	0.005662	0.006029	0.006885	0.007739
0.0	0.004429	0.004270	0.004712	0.007485	0.011697	0.009608
0.0	0.002419	0.001562	0.002122	0.002738	0.003614	0.003515
0.0	0.001233	0.001532	0.003536	0.002685	0.002917	0.003447
0.0	0.005840	0.003043	0.003741	0.005334	0.006845	0.006847
0.0	0.002276	0.001556	0.000893	0.002726	0.004608	0.003500
0.2	0.034945	0.042637	0.070772	0.074734	0.102439	0.095933
0.2	0.038642	0.044128	0.077429	0.077348	0.101663	0.099288
0.2	0.040292	0.043181	0.072470	0.075687	0.100947	0.097157
0.2	0.043042	0.041539	0.070000	0.072809	0.094983	0.093462
0.2	0.030227	0.041145	0.069220	0.072119	0.099686	0.092576
0.2	0.027873	0.039634	0.072144	0.069472	0.092323	0.089178
1.0	0.184855	0.224719	0.399046	0.393889	0.519319	0.505619
1.0	0.203213	0.215732	0.369567	0.378136	0.497636	0.485397
1.0	0.198748	0.219613	0.375057	0.384938	0.511100	0.494129
5.0	1.024497	1.044511	1.839075	1.830825	2.352620	2.350152
5.0	1.128297	1.123190	1.960976	1.968732	2.530950	2.527178
5.0	1.145603	1.140042	1.994428	1.998272	2.565619	2.565096

proach – normal distribution; $\boldsymbol{L}'_c = (0.00705, 0.01236, 0.01587)$; $\boldsymbol{L}'_d = (0.01150, 0.02016, 0.02588)$ modified Currie approach – Student's t-distribution (with $t_5 = 3.3649$); $\boldsymbol{L}'_D = (0.0131, 0.0230, 0.0295)$ Hubaux-Vos approach (using 98% confidence band).

These multivariate limits $\boldsymbol{L}_c$, $\boldsymbol{L}_d$ and $\boldsymbol{L}_D$ should lie within the predefined identification region $\boldsymbol{I}$ (e.g., should satisfy the fingerprint criterion, etc.) to confirm the presence of the analyte sought. One of the criteria used in the identification of unlabeled 2,3,7,8-TCDD is that the ratio (m/z at 320)/(m/z at 322) should lie within the interval (0.67, 0.87). The three limits $\boldsymbol{L}_c$, $\boldsymbol{L}_d$ and $\boldsymbol{L}_D$, given above, satisfy this criterion.

In practical applications, there is a tendency to use the simple point estimates x_d^{P} (Currie approach) or x_D^{P} (Hubaux and Vos approach) as the limit of detection. But if one wants to associate some confidence level with these DDLs, it is recommended to use x_d^{U} (Currie approach) or x_D^{U} (Hubaux and Vos approach) as the detection limit. From the above two tables, it is noticed that Hubaux-Vos approach generated slightly higher values of the DDLs. For hetroscedastic responses, it is recommended to use a separate regression model based upon data at low concentrations only for estimation of these DDLs.

Finally, an instrumental response $\boldsymbol{R}' = (r_1, r_2, \ldots, r_p)$ for which $\boldsymbol{R}$ belongs to the predefined identification region $\boldsymbol{I}$ and $r_i \geq L_{ic}$ (Currie approach), for all $i = 1, 2, \ldots, p$, indicates the presence of the analyte of interest, and a signal for which $r_i \geq L_{id}$ (Currie approach), or $r_i \geq L_{iD}$ (Hubaux-Vos approach) for all $i = 1, 2, \ldots, p$, represents a signal that is statistically significantly different from the blank signal confirming the detection of the analyte with high probability.

We acknowledge gratefully J. Armour, G.T. Flatman, N. Herron, D. Hewetson and J. Nocerino for their continuing help and valuable suggestions during the preparation of this manuscript. Although the research described in this article has been funded wholly by the U.S. Environmental Protection Agency through Contract 68-C0-0049 with Lockheed Engineering and Sciences Co., it has not been subjected to Agency review. Therefore, it does not necessarily reflect the view of the Agency.

REFERENCES

1 L.A. Currie, Anal. Chem., 40 (1968) 568.
2 A. Hubaux and G. Vox, Anal. Chem., 42 (1970) 849.
3 F.C. Garner and G.L. Robertson, Chemom. Intell. Lab. Syst., 3 (1988) 53.
4 I.T. Jolliffe, Principal Component Analysis, Springer Verlag, Heidelberg, 1986.
5 M.F. Delaney, Chemom. Intell. Lab. Syst., 3 (1988) 45.
6 R.A. Johnson and D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, 2nd edn., 1988.
7 N. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 2nd edn., 1981.
8 G.K. Shukla, Technometrics, 14 (1972) 547.
9 R.G. Krutchkoff, Technometrics, 11 (1969) 605.
10 E.J. Williams, Bull. Int. Statist. Inst., 43 (1969) 17.
11 H. Kaiser, Anal. Chem., 42 (1970) 26A.
12 IUPAC, Spectrochim. Acta, 33B (1978) 242.
13 J.C. Miller and J.N. Miller, Statistics for Analytical Chemistry, Ellis Horwood, London, 2nd edn., 1988.

Analytica Chimica Acta, 277 (1993) 215–221
Elsevier Science Publishers B.V., Amsterdam

Some graphical aids for univariate exploratory data analysis

Jiří Militký

Department of Textile Materials, Technical University, 46117 Liberec (Czech Republic)

Milan Meloun

Department of Analytical Chemistry, Technical University, 53210 Pardubice (Czech Republic)

(Received 30th June 1992)

Abstract

The main parts of exploratory data analysis (EDA) are discussed. For data presentation the quantile plot and quantile-box plot are proposed. Special techniques for empirical probability density construction and empirical quantile–quantile plot creation are described. Some graphically oriented methods for selection of optimum power transformations are presented. These graphical aids in EDA are demonstrated on Hinkley's well known data.

Keywords: Exploratory data analysis; Power transformation; Probability density; Univariate data analysis

The classical approach to statistical data analysis is based on some strong assumptions about their statistical nature such as independence, normality and homogeneity. Frequently the data are less than ideal and their effective analysis can be realized in two stages. The first stage is exploratory data analysis (EDA), where data for uncovering typical relationships and patterns are surveyed and treated. The second stage is confirmatory data analysis (CDA), where probability models are created and tested.

According to Tukey [1], EDA is "detective work". It uses as its tools various descriptive and graphically oriented techniques that are typically free from strict statistical assumptions about data. These techniques are often called "distribution free" and are based only on basic assumptions such as continuity and differentiability of underlying density. EDA techniques are especially effective for the investigation of the statistical behaviour of data from new or non-standard measurements or for the creation of probability models. A typical chemometric example is trace analysis.

One of most frequent tasks in statistical data analysis is the one-sample problem based on a sample $x_1, \ldots, x_n$ representing the behaviour of a univariate (random) variable x. For this case EDA has three main goals: visualization of statistical features of the sample; construction of an empirical sample distribution and comparison of this distribution with theoretical ones; and data transformation for improving their distribution such as symmetrizing or normalization.

The realization of these goals involves the utilization of techniques well suited especially for

Correspondence to: J. Militký, Department of Textile Materials, Technical University, 46117 Liberec (Czech Republic).

small and moderate sample sizes [2]. The most popular EDA methods for one-sample problems with applications in chemistry have been surveyed [3].

In this paper, selected simple EDA techniques are discussed. The full set of EDA analysis techniques used in the module Basic Statistics in the package ADSTAT has been described elsewhere [3]. Some EDA techniques are demonstrated on Hinkley's well known data (sample of 30 values) [4].

SOME BASIC CONCEPTS

The EDA techniques for small and moderate samples are based on order statistics:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

which are the sample values (assumed to be distinct) arranged in increasing order. Let $F_e(x)$ be the distribution function from which values x_i are sampled. It is well known that the transformed random variable

$$Z_{(i)} = F_e[x_{(i)}] \tag{1}$$

has independently of the distribution function F_e the beta distribution $Be\ [i, n-i+1]$. Its mean value is given by

$$E[Z_{(i)}] = i/(n+1) \tag{2}$$

The elements V_{ij} of the covariance matrix V for all $Z_{(i)}(i = 1, \ldots, n)$ are simple functions of i, j and n only. Using back transformations of $E[Z_{(i)}]$, we obtain the relationship

$$E[x_{(i)}] = F_e^{-1}[Z_{(i)}] = Q_e(P_i) \tag{3}$$

where $Q_e(P_i)$ is a quantile function and $P_i = i/(n+1)$ is the cumulative probability. A detailed description of the quantile function properties and its advantages for constructing empirical sample distributions was given by Parzen [5].

From Eqn. 3 we can deduce that the order statistic $x_{(i)}$ is a raw estimate of the quantile function $Q_e(P_i)$ in the position of P_i. For estimation of the quantile $x_P \equiv Q_e(P)$ at a value $i/(n+1) < P < (i+1)/(n+1)$, the piecewise linear interpolation

$$x_P = (n+1)(P - i/(n+1))[x_{(i+1)} - x_{(i)}] + x_{(i)} \tag{4}$$

can be used. The variance $D(x_P)$ can be calculated from the equation

$$D(x_P) = P(1-P)/[nf_e^2(x_P)] \tag{5}$$

where $f_e(x_P)$ is a probability density function corresponding to the distribution function F_e.

Equation 4 can be used for estimation of sample quantiles x_{P_i} or x_{1-P_i} for $P_i = 2^{-i}$ $(i = 1, \ldots, n)$. These quantiles are called letter values [6]. All letter values except $i = 1$ (median) are in pairs. For example, we can estimate the lower quartile $x_{0.25}$ $(P_i = 0.25)$ and upper quartile $x_{0.75}$ $(P_i = 0.75)$, etc.

For EDA purposes the modified definition of cumulative probability

$$P_i = (i - 0.375)/(n + 0.25) \tag{6}$$

is often used. A discussion of a suitable definition of P_i was presented by Looney and Gulledge [7].

TECHNIQUES FOR DATA PRESENTATION

For the graphical representation of data, many simple techniques such as the stem-leaf plot, box plot, dot plot [1] and digdot plot [8] have been proposed. The quantile-box plot (QBP) and quantile plot (QP) are selected here. Symmetry and tail length can be characterized by use of the g–h distribution system [3,10].

Quantile plot

An empirical sample quantile plot $Q(P)$ is constructed as a dependence of $x_{(i)}$ on P_i. From patterns of points some statistical features of data such as symmetry, local concentration and rough normality can be simply recovered. A detailed interpretation of QP has been given elsewhere [9].

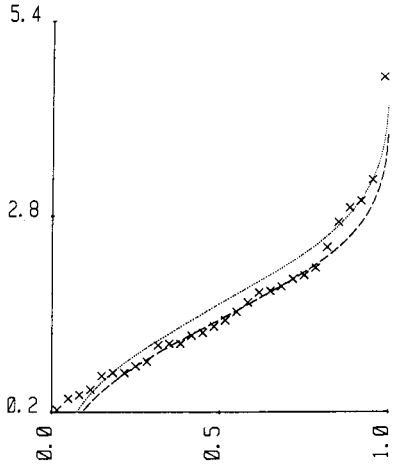


Fig. 1. Quantile plot for Hinkley's data. ———, Robust estimate; ·····, classical estimate.

For comparative purposes, the quantile functions of a normal distribution

$$Q_N(P) = \mu + \sigma u_P \tag{7}$$

where u_P are quantiles of the standard normal distribution $N(0, 1)$ and μ and σ are estimators of location and scale, are superimposed on QP.

Two different normal quantile functions are plotted. One is based on the sample mean x_M and sample standard deviation s. The other uses robust quantile estimators $\mu = x_{0.5}$ and $\sigma = (x_{0.75} - x_{0.25})/1.349$.

Figure 1 shows the quantile plot for Hinkley's data. It is clear that these data are positively skewed and can be approximated in the middle part by a normal distribution (with robust estimators of location and scale).

Quantile-box plot

A quantile-box plot (QBP), proposed by Parzen [5], is an extension of the idea of a box plot introduced by Tukey [1]. A QBP consists of quantile function graph (see previous section) on which various boxes with vertices

$$[(P, Q(P)); (P, Q(1-P)); (1-P, Q(P)); (1-P, Q(1-P))]$$

are superposed. One usually chooses quartile ($P = 1/4$), octile ($P = 1/8$) and sedecile ($P = 1/16$) boxes.

TABLE 1

Theoretical tail lengths

Distribution	T_3	T_4
Normal	0.534	0.822
Rectangular	0.405	0.559
Laplacian	0.693	1.098

Within the quartile box one draws a median line with the vertices

$$[(0.25, Q(0.5)), (0.75, Q(0.5))]$$

An approximate confidence interval for the median $x_{0.5} \equiv Q(0.5)$ is indicated by a vertical line with vertices

$$[(0.5, x_{0.5} \pm 1.57(x_{0.75} - x_{0.25})/\sigma)]$$

The QBP with some quantile measures of symmetry such as [5]

$$S_i = [x_{0.5} - 0.5(x_{P_i} + x_{1-P_i})]/(x_{1-P_i} - x_{P_i}) \tag{8}$$

($i = 2, 3, 4$) and tail length [5]

$$T_i = \ln[(x_{1-P_i} - x_{P_i})/(x_{0.75} - x_{0.25})] \tag{9}$$

($i = 3, 4$) can be used for the description of data peculiarities at various distances from median. The theoretical T_i values for octiles ($i = 3$) and sedeciles ($i = 4$) are presented in Table 1.

As described [5], the QBP can also be used for the identification of polymodal distributions and outliers.

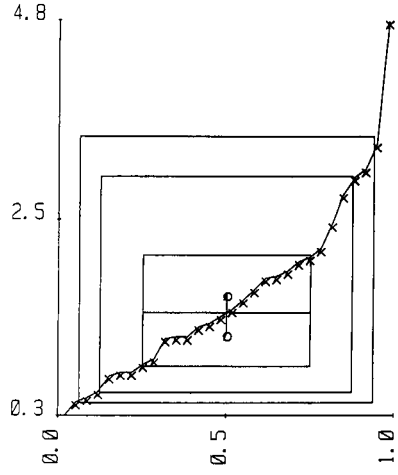


Fig. 2. Quatile-box plot for Hinkley's data.

Figure 2 shows the QBP for Hinkley's data. Corresponding values of S_i and T_i are presented in Table 2. It is evident that the data are positively skewed and contain one outlying observation.

CONSTRUCTION AND COMPARISON OF SAMPLE DISTRIBUTION

As an estimator of the empirical probability density function, histograms with variable bins and kernel type density are constructed. Comparison of the sample distribution with theoretical distribution is based on variants of the Quantile–Quantile (Q–Q) plot. The probability–probability and transformed distribution function plots [5] can also be used [11].

Empirical probability density

A histogram is a piecewise constant estimator of sample probability density (PDF). The histogram height in the jth class bounded by values (t_{j-1}, t_j) is calculated from the relationship

$$f_H(x) = (nh_j)^{-1} C_n(t_{j-1}, t_j) \qquad (10)$$

where the function $C_n(a, b)$ denote the number of sample elements within $\langle a, b\rangle$ and $h_j = t_j - t_{j-1}$ is the length of the jth interval. Now, the problem encountered is the choice of boundary values $\{t_j\}(j = 1, \ldots, M)$, the number of class intervals M and their lengths h_j with respect to the histogram quality. In our programs the simple data-based two-stage technique is used.

In the first stage, the number of class intervals

$$M_0 \approx \text{int}\left[2.46(n-1)^{0.4}\right]$$

is assessed. In the second stage, the individual lengths h_j are determined. The estimation of h_j is based on the requirement of equal probability in all classes. For this purpose the empirical quantile function $Q(P)$ based on order statistics $x_{(i)}$ is implemented.

In practice, the P-axis is divided into identical intervals of size $1/M_0$. For these intervals the corresponding quantile estimates $t_j = x(j/M_0)$ are constructed by using Eqn. 4, where $P = j/M_0$. Practical experiences has hitherto demonstrated that this construction is suitable even for strongly skewed sample distributions. Figure 3 shows the histogram with a normal probability density superimposed for Hinkley's data.

The kernel-type non-parametric estimator of sample probability density $f(x)$ can be constructed on the basis of the Lejenne–Dodge–Kaelin procedure [12].

TABLE 2

Characteristics of symmetry and tail length for Hinkley's data

Quantile	i	S_i	T_i
Quartile	2	−0.025	0.000
Octile	3	−0.134	0.712
Sedecile	4	−0.163	0.879

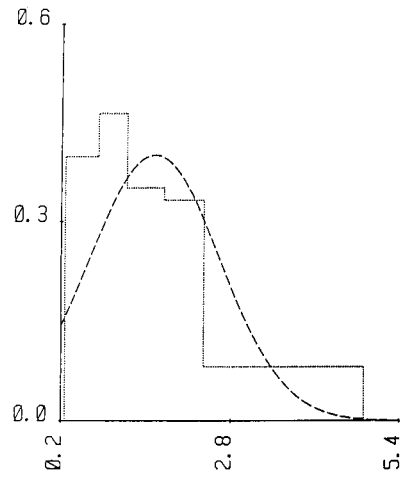


Fig. 3. Histogram for Hinkley's data. ———, Normal PDF; ·····, histogram.

Comparison of sample distribution

For the purpose of comparison of empirical sample distributions with theoretical distributions, variants of the Q–Q plot are suitable. The classical Q–Q plot is based on a comparison of the empirical quantile function $Q(P_i) \equiv x_{(i)}$ with a chosen theoretical quantile function $Q_T(P_i)$. For theoretical distribution functions of the type $F_T[(x-\mu)/\sigma]$ it is advantageous to use the standardized quantile function $Q_{TS}(P_i)$.

When the empirical and theoretical distributions coincide, the relationship

$$x_{(i)} = \mu + \sigma Q_{TS}(P_i) \tag{11}$$

is valid. Here usually μ is the shape parameter and σ is the parameter of scale. For some three-parameter distribution the shape factor is usually a parameter of the plot. Our programs select a shape factor value that straightens the individual points best.

Owing to the strong dependence among order statistics and their non-constant variance, the Q–Q plot gives a very patterned appearance and the degree of linearity is difficult to quantify. Michael [13] introduced the stabilized probability plot and Kafander and Spiegelman [14] proposed the conditional Q–Q plot.

For EDA purposes we use the empirical probability plot (EPP) [2]. In EPP the quantiles $Q_{TS}(P_i)$ are replaced with simulated ones T_i generated from chosen theoretical distribution.

The process of the computation of T_i can be divided into three main steps: from an assumed theoretical distribution the simulated samples $\{x_i^j\}(i = 1, \ldots, n;\ j = 1, \ldots, 25)$ are generated; from all samples $(j = 1, \ldots, 25)$ the order statistics $x_{(i)}^j$ are computed; and the simulated quantiles T_i are medians from corresponding order statistics of all simulated samples:

$$T_i = \text{med}\{x_{(i)}^1, \ldots, x_{(i)}^{25}\}$$

Based on the second largest values and second lowest values in the sequence $\{x_{(i)}^1, \ldots, x_{(i)}^{25}\}$, the boundary of the 85% confidence intervals can be constructed. An analogous procedure for the case of logistic regression has been described [15].

Figure 4 shows the classical Q–Q plot and Fig. 5 shows the EPP plot for Hinkley's data. In both plots the normal distribution is selected as a theoretical distribution. The systematic deviation from linearity indicated a non-normal sample.

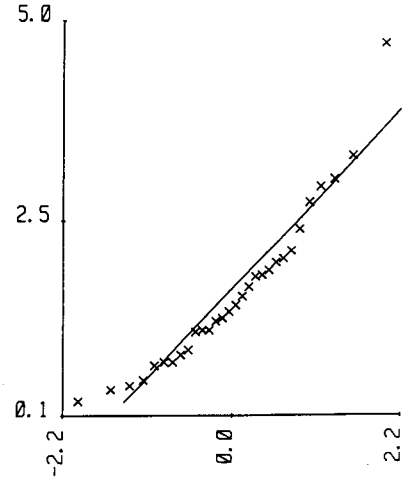


Fig. 4. Q–Q plot for Hinkley's data (theoretical is normal distribution).

POWER TRANSFORMATION OF DATA

Power transformation is used in the context of EDA as a tool for simplifying the data distribution. Suitable power–law transformations may result in a distribution that is nearly symmetrical and perhaps more nearly normal.

In many instances the symmetrizing of the data distribution by using a simple power transformation:

$$\begin{aligned} y = g(x) &= x\hat{}(\beta) && \text{for } \beta > 0 \\ y = g(x) &= \ln x && \text{for } \beta = 0 \\ y = g(x) &= -x\hat{}(-\beta) && \text{for } \beta < 0 \end{aligned} \tag{12}$$

can be obtained. This transformation is not scale invariant and is not a continuous function of β. It

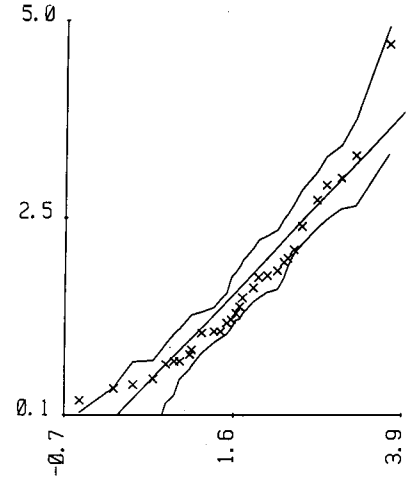


Fig. 5. EPP plot for Hinkley's data (theoretical is normal distribution).

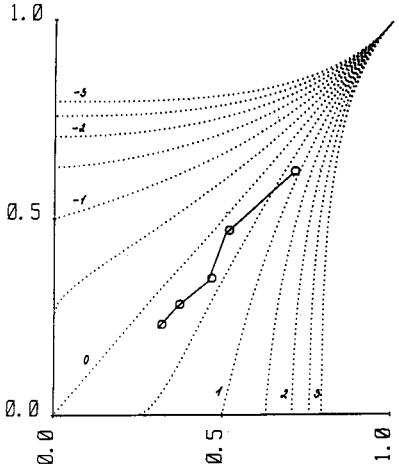


Fig. 6. Selection graph for Hinkley's data.

is suitable for positive data only. Optimum transformation can be selected by minimizing some robust measures of skewness:

$$g_R = [(y_{0.75} - y_{0.5}) - (y_{0.5} - y_{0.25})] / (y_{0.75} - y_{0.25}) \quad (13)$$

As a diagnostic tool a selection graph can be simply constructed. This graph is based on the requirement of symmetry of quantiles about the median. This requirement can be mathematically described by the relationship [16]

$$(x_{P_i}/x_{0.5})^{\wedge}(\beta) + (x_{0.5}/x_{1-P_i})^{\wedge}(-\beta) = 2 \quad (14)$$

Letter values, where $P_i = 2^{-i}$ for $i = 2, 3, 4, \ldots$, are usually chosen. The selection graph has on the y-axis the quantities $x_{P_i}/x_{0.5}$ and on the x-axis the quantities $x_{0.5}/x_{1-P_i}$. For comparison of computed points with ideal courses for constant β, the solution of the equation

$$y^{\wedge}(\beta) + x^{\wedge}(-\beta) = 2 \quad 0 \le x \le 1;\ 0 \le y \le 1 \quad (15)$$

is superimposed on the graph. Figure 6 shows the selection graph for Hinkley's data. It is clear that it is suitable to choose $\beta \approx 1/3$. The value $\beta = 0.2$ leads to the minimum value $g_R = 0.06$.

Another exploratory technique for graphical estimation of power β was described by Emerson and Stoto [17].

The Box–Cox power transformation family, which is a continuous function of β, can be defined by

$$y = g(x) = [x^{\wedge}(\beta) - 1]/\beta \quad \text{for } \beta \neq 0$$
$$y = g(x) = \ln x \quad \text{for } \beta = 0 \quad (16)$$

This transformation can be used for positive data only. After slight modification the range of applicability can be arbitrarily extended.

The properties of this transformation family have been studied in depth (e.g., [6]). Based on the assumption that for some β, y is a normally distributed variable $N(\mu_y, \sigma_y^2)$, the likelihood function can be constructed. The logarithm of likelihood function has the form

$$\ln L(\beta) = -n/2 \ln(s_y^2) + (\beta - 1) \sum_i \ln x_i \quad (17)$$

where s_y^2 is the sample variance of transformed data. The likelihood function can be plotted against β in a suitable range (the standard range is $-3 \le \beta \le 3$). On this plot the $100(1-\alpha)\%$ confidence interval of power β:

$$2[\ln L(\beta^*) - \ln L(\beta)] \le \chi^2(1) \quad (18)$$

is superimposed, where β^* is the maximum likelihood estimator of β. In the confidence interval defined by Eqn. 18 are all values of β for which $\ln L(\beta) \in \ln L(\beta^*) - 0.5\chi^2(1)$, where χ^2 is a quantile of the χ-squared distribution. From the

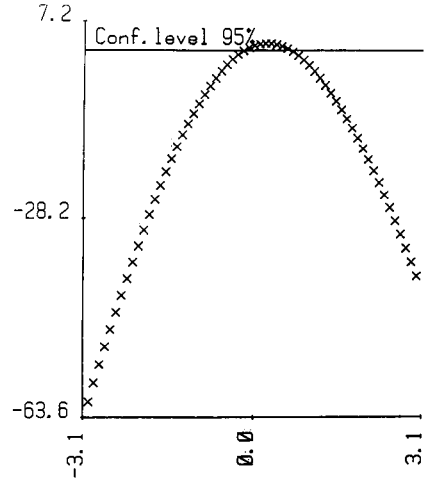


Fig. 7. Likelihood function plot for Hinkley's data.

width of the confidence interval the quality of power transformation can be indicated.

Figure 7 shows the likelihood function plot for Hinkley's data. The optimum power maximizing ln $L(\beta)$ is $\beta = 0.2$.

The quality of power transformation can also be described by using the above-discussed graphical techniques.

PROGRAM SYSTEM ADSTAT

ADSTAT contains eight modules of statistical methods for univariate and multivariate data [3]. Manipulations with ADSTAT are very simple by using pull-down menus and panes. Individual program modules are built in a window-like environment. This environment includes the powerful block-oriented data editor, context-sensitive help system and unified graphical presentation. Exploratory methods included in the module Basic Statistics can be divided to three main parts: techniques for presentation of data; construction of empirical sample distribution and its comparison with twelve theoretical distributions; and power transformation of data by using Eqns. 12 and 16.

The above-mentioned and more complex EDA techniques described elsewhere [3] are used.

Conclusion

The program system ADSTAT is well suited for EDA of one-sample problems on personal computers. An extensive description of algorithms used in ADSTAT and examples of its utilization for the analysis of chemical data are available [3].

REFERENCES

1 J.W. Tukey, Exploratory Data Analysis, Addison Wesley, Reading, MA, 1977.
2 J. Militký, presented at COMPSTAT '84 Conference, Prague, 1984.
3 M. Meloun, J. Militký and M. Forina, Chemometrics for Analytical Chemistry, Horwood, Chichester, 1992.
4 D.V. Hinkley, Appl. Stat., 26 (1977) 67.
5 E. Parzen, J. Am. Stat. Assoc., 74 (1985) 105.
6 D.C. Hoaglin, F. Mosteler and J.W. Tukey (Eds.), Understanding Robust and Exploratory Data Analysis, Wiley, New York, 1983.
7 S.W. Looney and T.R. Gulledge, Staistician, 34 (1985) 297.
8 S. Hunter, Am. Stat., 42 (1988) 54.
9 J. Chambers, W.S. Cleveland, B. Kleiner and P.A. Tukey, Graphical Methods for Data Analysis, Duxbury Press, Boston, 1983.
10 D.C. Hoaglin, F. Mosteler and J.W. Tukey (Eds.), Exploring Data, Tables, Trends and Shapes, Wiley, New York, 1985.
11 J. Militký, presented at COMPSTAT '88 Conference, Copenhagen, 1988.
12 M. Lejenne, Y. Dodge and E. Kaelin, presented at COMPSTAT '82 Conference, Toulouse, 1982.
13 J.R. Michael, Biometrika, 70 (1983) 11.
14 K. Kafander and C.H. Spiegelman, Comput. Stat. Data Anal., 4 (1986) 167.
15 J.M. Landwehr, D. Pregibon and A.C. Shoemaker, J. Am. Stat. Assoc., 79 (1984) 61.
16 W.G.S. Hines and R.J.H. Hines, Am. Stat., 41 (1987) 21.
17 J.D. Emerson and M.A. Stoto, J. Am. Stat. Assoc., 77 (1982) 103.

Analytica Chimica Acta, 277 (1993) 223–238
Elsevier Science Publishers B.V., Amsterdam

Developing models for infinite dilution activity coefficients using factor analysis methods

Russell B. Poe and Sarah C. Rutan

Department of Chemistry, Virginia Commonwealth University, Richmond, VA 23284-2006 (USA)

Mitchell J. Hait and Charles A. Eckert

School of Chemical Engineering and Chemistry, Specialty Separations Center, Georgia Institute of Technology, Atlanta, GA 30332 (USA)

Peter W. Carr

Department of Chemistry, University of Minnesota, Minneapolis, MN 55455 (USA)

(Received 3rd July 1992)

Abstract

This paper illustrates a novel use of factor analysis to characterize models for infinite dilution activity coefficients. Different forms of a modified cohesive energy density (MOSCED) equation were used to create synthetic data matrices to be analyzed by factor analysis. The abstract factors from these matrices were used to target test the different model parameters. By using the synthetic data, the factors contributing to the different forms of the MOSCED equation that were most susceptible to noise were identified. Factor analysis and target testing were then used to characterize a database of experimental infinite dilution activity coefficients for 29 solvents and 31 solutes. The factor analysis results for one form of the MOSCED equation were used to evaluate different physicochemical scales for dispersion, dipolarity and hydrogen bonding interactions.

Keywords: Principal component analysis; Database; Factor analysis; Infinite dilution activity coefficients

In recent years, there has been a great deal of attention given to measuring and predicting infinite dilution activity coefficients. This fundamental area of chemistry is extremely important in developing new production processes and refining existing ones in the chemical industry [1]. Infinite dilution activity coefficients can be used in predicting the results of liquid–liquid extractions [2], chromatographic retention times [3], and multicomponent vapor–liquid equilibria for completely miscible systems [4]. However, accurately measuring infinite dilution activity coefficients is an extremely difficult and tedious task. There are many methods for measuring infinite dilution activity coefficients, including head space gas chromatography [5], ebulliometry [6] and dynamic gas chromatography [7]; each method has its own advantages and limitations. Most of these methods suffer from several potential errors, and it is difficult to obtain reliable results without very painstaking experimental techniques. It is neither practical nor possible to measure infinite dilution

Correspondence to: S.C. Rutan, Department of Chemistry, Virginia Commonwealth University, Richmond, VA 23284-2006 (USA).

activity coefficients for all possible combinations of solvents and solutes; therefore, it would be advantageous to be able to accurately predict infinite dilution activity coefficients from pure component properties.

There are several methods available for predicting infinite dilution activity coefficients (γ^{∞}), the leading methods being the UNIFAC (UNIQUAC functional group activity coefficient) [8,9] and MOSCED (modified separation of cohesive energy density) [9–11] models. UNIFAC is a method based on summing group contributions of functional groups present in the solute and solvent, and MOSCED is an extension of regular solution theory where the mixture properties are determined from the different physical interactions: dispersion, induction, dipolar interactions, and hydrogen bonding.

One of the advantages of MOSCED is that the predictions of infinite dilution activity coefficients are made based only on pure component properties. That is, binary parameters specific to the system are not needed. However, there is a need to determine appropriate scales to describe these pure component properties for representing dispersion, induction, dipolar, and hydrogen bonding interactions.

Over the past few years, factor analysis methods have been widely used to characterize data for a variety of linear systems. Matrices consisting of data as a function of chromatographic retention time and UV–visible absorption wavelength have been successfully analyzed using factor analysis, however, data that may be modeled by more complicated equations are normally not studied using factor analysis. (In this study, the terms factor analysis, eigenanalysis and principal component analysis (PCA) will be used interchangeably.) In our work, we illustrate the use of factor analysis to characterize a database of infinite dilution activity coefficients. In addition, using PCA and target testing, we are able to evaluate different models for predicting infinite dilution activity coefficients. The PCA procedure yields a set of principal components which describe the major sources of variation in a data set. Target testing allows potentially significant pure component properties to be evaluated, and allows comparisons of different model formulations. We have generated synthetic data based on three different formulations of a MOSCED type model. Analysis of the synthetic data enables us to determine the number of linearly independent factors that are required to describe the different types of MOSCED models. The relative importance of each of the terms, and their sensitivity to the presence of experimental errors is also studied.

THEORY

The List of symbols contains a summary of the notation used in this paper and a listing of all parameters used.

The database used in this work consisted of a data matrix of infinite dilution activity coefficients for 29 solvents and 31 solutes, along with a set of physical properties for these solvents and solutes. A more detailed description of the database is presented in the Experimental section of the paper.

Data pretreatment

The database of infinite dilution activity coefficients was examined and any row or column that contained more than 30% missing values was not included in the matrix to be analyzed. The data matrix that remained had 29 rows (solvents) and 31 columns (solutes). Any missing infinite dilution activity coefficients were then replaced with estimated values obtained from UNIFAC or MOSCED to complete the matrix. These values were thought to be better approximations of the missing values than the mean of the column or row and should contribute some chemical information. However, the average error of these estimated infinite dilution activity coefficients is higher than that of the experimentally determined values and it is certainly desirable to replace these initial estimates with values that are more consistent with the experimentally determined data.

Missing data routine

After the original missing values are estimated by UNIFAC or MOSCED, the values were improved using a factor analysis approach for com-

pensating for missing values [12]. The logarithms of the activity coefficients are used to form a matrix **D**. The matrix **D** can be expressed in the terms of its principal components

$$\mathbf{D} = \mathbf{USV}^{T} \quad (1)$$

where **U** is the normalized row-factor matrix, **S** is a diagonal matrix composed of the square roots of the eigenvalues and $\mathbf{V}^{T}$ is the normalized column-factor space. Next, the number of primary eigenvalues (rank) is estimated using the IND function [13], the residual standard deviation (RSD) [14] and the *F*-test [15]. The primary eigenvalues are assumed to describe the variance resulting from physical and chemical interactions contained within the data matrix **D** and the secondary eigenvalues are assumed to describe the variance of the noise. The IND function, developed by Malinowski [13], is an empirical calculation using the eigenvalues that has been shown to reach a minimum value when the correct number of factors is chosen. The RSD and the *F*-test methods are based on the fact that the secondary eigenvalues should reflect the variance of the noise affecting the measurements. Once the rank of the data matrix is determined, the data matrix is regenerated using only the primary factors

$$\overline{\mathbf{D}} = \overline{\mathbf{U}}\,\overline{\mathbf{S}}\overline{\mathbf{V}}^{T} \quad (2)$$

The reduced matrix $\overline{\mathbf{D}}$ contains less noise than the original matrix **D**. The missing values regenerated from the reduced matrix should be more consistent with the experimental values than the original MOSCED or UNIFAC values. These estimated values are then substituted for the UNIFAC and MOSCED values in the original matrix, giving a new matrix $\mathbf{D}_2$. Principal component analysis (PCA) is now performed on this matrix, using the same number of primary eigenvectors as in the first iteration. The missing values are again replaced with the reduced factor regenerated values. This process proceeds until the sum of squares of the deviations in the adjusted values in succeeding iterations decreases to a value below a preset tolerance. The final matrix contains the original experimental data and the iteratively modified MOSCED and UNIFAC values. Figure 1 is a flow chart for this missing data routine. The solid triangles represent experimental values, the circles represent the missing values that were estimated from UNIFAC or MOSCED, and the hollow triangles are estimates for the missing values obtained after repeated iterations.

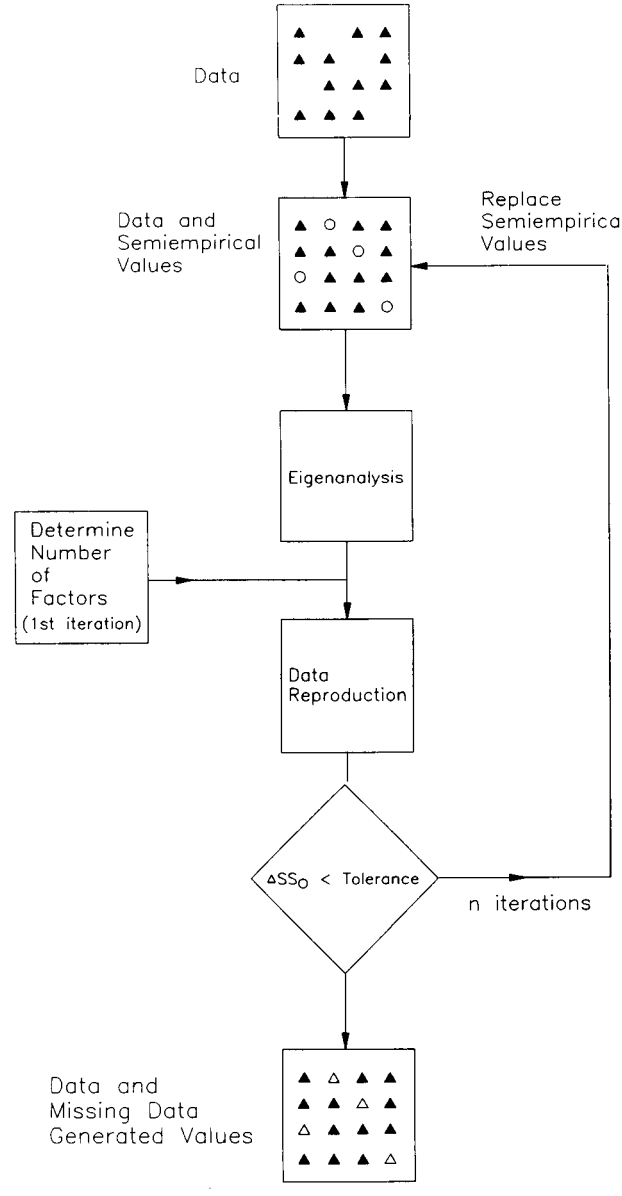


Fig. 1. Flow chart for missing data factor analysis program. (▲) Experimental value, (○) missing values that were estimated from UNIFAC or MOSCED, and (△) estimate for the missing value obtained after repeated iterations.

Principal component analysis is performed one final time on the complete data matrix, including the values returned by the missing data routine. The normalized row-factors and column-factors are used to construct cluster diagrams and to correlate with the solute and solvent properties.

Simulated alkane data

Before using principal component analysis on the data matrix of experimentally determined infinite dilution activity coefficients, several synthetic data matrices were constructed to evaluate the use of PCA in studying this type of data, since most of the studies reported on PCA have been for systems with relatively simple linear models [16].

The model used to construct the first synthetic data matrix follows regular solution theory [17], which is expected to hold for non-polar solutes dissolved in non-polar solvents, and is expressed as follows

$$\ln \gamma_2^{\infty} = V_2/RT(\delta_1 - \delta_2)^2 + 0.3 \ln(V_2/V_1) \quad (3)$$

where V is the volume, δ is the Hildebrand solubility parameter, R is the gas constant and T is the absolute temperature. The subscript 1 denotes a solvent parameter and the subscript 2 denotes a solute parameter. The first term comes from regular solution theory and the second term is an approximation of a Flory–Huggins type term to account for differences in molecular size [10]. The 0.3 coefficient was derived from fits of an equation with this form to data for alkane solutes and solvents with moderate variation in size [18]. This format was found to be much better at describing the correct trends and signs in the $\ln \gamma^{\infty}$ data than the original Flory–Huggins term used for MOSCED $\{1 - (V_2/V_1)^{aa} + \ln(V_2/V_1)^{aa}$ [10]}.

For PCA, it is assumed that the data in a matrix **D** can be represented in a bilinear form

$$\mathbf{D} = \sum_{k=1}^{n} s_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}} \quad (4)$$

where n is the number of linear factors contributing to the variance in the data, the $\mathbf{u}_k$ are solvent dependent factors, the $\mathbf{v}_k$ are solute dependent factors, and the s_k are scaling factors. To find the rank of the synthetic data, Eqn. 3 must be factored into a form consistent with Eqn. 4, as shown in the following equation

$$\ln \gamma_2^{\infty} = V_2\delta_1^2/RT - 2V_2\delta_2\delta_1/RT + V_2\delta_2^2/RT + 0.3 \ln V_2 - 0.3 \ln V_1 \quad (5)$$

Analysis of this equation indicates that there are four solvent dependent terms ($\mathbf{u}_1 - \mathbf{u}_4$) and four solute dependent terms ($\mathbf{v}_1 - \mathbf{v}_4$). Table 1 lists the four terms for the solutes and for the solvents. This analysis suggests, for example, that a target test (i.e., linear regression) of $\ln V_1$ to the four solvent factors ($\mathbf{u}_1 - \mathbf{u}_4$) should yield a good fit.

TABLE 1

Linear factors for alkane model

Solute	Solvent
V_2	δ_1^2
$V_2\delta_2$	δ_1
$V_2\delta_2^2/RT + 0.3 \ln V_2$	Offset_1
Offset_2	$\ln V_1$

Simulated MOSCED data

In an attempt to more accurately model activity coefficients, a more versatile and comprehensive model must be used than that described above. The more complex model should consider non-polar, polar, and hydrogen bonding interactions; in this case a simplified form of the original MOSCED model was used

$$\ln \gamma_2^{\infty} = V_2/RT\left[(\lambda_1 - \lambda_2)^2 + q_1^2q_2^2(\tau_1 - \tau_2)^2 + (\alpha_1 - \alpha_2)(\beta_1 - \beta_2)\right] + 0.3 \ln(V_2/V_1) \quad (6)$$

where λ is a dispersion parameter, q is an induction parameter, τ is a dipolarity parameter, α is a hydrogen bond acidity parameter, β is a hydrogen bond basicity parameter, and the remaining parameters are the same as for Eqn. 3. This equation is similar to the equation described by Thomas and Eckert [10]; however the Flory–Huggins term was simplified based on fits of alkane solute and solvent data, as mentioned above, and the asymmetry correction terms were left out of

TABLE 2

Linear factors for MOSCED model

Solute	Solvent
V_2	$\lambda_1^2 + \alpha_1\beta_1$
$V_2\lambda_2$	λ_1
$V_2q_2^2$	$q_1^2\tau_1^2$
$V_2q_2^2\tau_2$	$q_1^2\tau_1$
$V_2q_2^2\tau_2^2$	q_1^2
$V_2\alpha_2$	β_1
$V_2\beta_2$	α_1
$V_2(\lambda_2^2 + \alpha_2\beta_2)/RT + 0.3 \ln V_2$	Offset_1
Offset_2	$\ln V_1$

the model. The asymmetry correction terms were included in the original MOSCED model to account for the fact that the γ^∞ for solute A in solvent B is not normally the same as the γ^∞ for solute B in solvent A which is predicted by the regular solution theory formalism (except for differences in molar volume, which are usually small). Current modifications to the MOSCED model have defined separate scales for solute and solvent parameters that removes the need for these asymmetry correction factors [19]. When the MOSCED equation is factored and expressed in the form of Eqn. 4, there are 9 terms contributing to the variance of the data. Table 2 lists the terms for both the solvents and solutes.

Simulated modified MOSCED data

An alternative model for MOSCED was tested that did not use the q^2 parameters for modeling inductive interactions but instead used the polarizability parameter used for modeling dispersion interactions to describe the induced dipole interactions. This alternative model is

$$\ln \gamma_2^\infty = \frac{V_2}{RT}\left[(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_2)(\tau_1 - \tau_2) + (\tau_1 - \tau_2)^2 + (\alpha_1 - \alpha_2)(\beta_1 - \beta_2)\right] + 0.3 \ln(V_2/V_1) \quad (7)$$

This model has a rank of seven; the terms are listed in Table 3.

TABLE 3

Linear factors for modified MOSCED model

Solute	Solvent
V_2	$\lambda_1^2 + \lambda_1\tau_1 + \tau_1^2 + \alpha_1\beta_1$
$V_2(2\lambda_2 + \tau_2)$	λ_1
$V_2(\lambda_2^2 + \lambda_2\tau_2 + \tau_2^2 + \alpha_2\beta_2)/RT + 0.3 \ln V_2$	Offset_1
$V_2(\lambda_2 + 2\tau_2)$	τ_1
$V_2\beta_2$	α_1
$V_2\alpha_2$	β_1
Offset_2	$\ln V_1$

EXPERIMENTAL

Principal component analysis was used in the analysis of a matrix that contained 685 experimentally determined infinite dilution activity coefficients. The values for this matrix were obtained from a large database compiled by Hait et al. [19], which contained infinite dilution activity coefficients obtained using several different methods, each with a different level of experimental uncertainty and accuracy. Where no experimental values were available, the database contained several estimates from different versions of UNIFAC and MOSCED. The data matrix that was analyzed here contained γ^∞ values for 29 solvents and 31 solutes, therefore 899 values are required to complete the matrix to use principal component analysis. The final data matrix of infinite dilution activity coefficients contained 75% experimentally determined values, 20% of the values estimated by UNIFAC and 5% of the values estimated by MOSCED. A PCA program which predicts values for the missing entries (estimated initially by the UNIFAC or MOSCED values) was used to generate a complete data matrix for subsequent data analysis, as described in the Theory section (see Fig. 1). In addition to infinite dilution activity coefficients, the database contained various solvatochromic and physical parameters for the pure components [20–26].

Simulated data were created by generating a vector of random numbers for each parameter, over a physically realistic range of values. For all simulated models, data were generated for 29 solvents and 31 solutes, as these were the dimen-

TABLE 4

Ranges of synthetic parameters

Solute parameter	Range	Solvent parameter	Range
V_2	50–120	V_1	50–200
δ_2	7–15	δ_1	7–15
λ_2	7–10	λ_1	7–10
q_2	0.9–1.0	q_1	0.9–1.0
τ_2	0–6	τ_1	0–6
α_2	0–10	α_1	0–10
β_2	0–10	β_1	0–10

sions of the experimental ln γ^∞ matrix. The ranges for the parameters used to simulate data according to the models given in Eqns. 3, 6, and 7 are listed in Table 4. These parameters were saved for target testing of the resulting synthetic data matrices.

All computations were performed on a 16-MHz 80386 computer equipped with an 80387 coprocessor, and running under the MS-DOS operating system. The principal components analysis, missing data and target testing programs were written using Turbo Pascal, version 6.0 (Borland).

RESULTS AND DISCUSSION

Effect of normally distributed noise on synthetic data

Using the alkane model described by Eqn. 3, three matrices were generated with identical values used as input parameters; however, Gaussian distributed noise with a standard deviation of 0.05 or 0.20 ln γ^∞ units was added to the activity coefficient values for two of the matrices. This resulted in three matrices of simulated ln γ^∞ values with approximately 0, 5 and 20% error in γ^∞. The three matrices were analyzed by PCA and the abstract factors saved for target testing. The terms in Table 1 were used as test vectors to probe the effects of normally distributed noise on the performance of PCA for this type of model.

After PCA of the data matrices containing 0 and 5% noise, the IND function, RSD and *F*-test indicated that there were 4 significant factors that contributed to the data; however, these tests showed only three significant factors for the matrix containing 20% noise. This indicated that some of the model variations were indistinguishable from the noise, or that the four factors were no longer linearly independent, within the variance of the noise. This result is important because it indicates that important chemical interactions in the real data may be obscured by moderate noise levels, and that to adequately understand these interactions, the noise contributions to the real data should be substantially less than 20%.

We explored several means for evaluating the results or quality of fit obtained from the target testing methodology. The SPOIL function is one method of determining the quality of fit [16], however, we found this parameter not as useful as the normalized standard error (NSE), defined as

$$\mathrm{NSE} = \sqrt{\frac{\dfrac{\sum (x_i - \hat{x}_i)^2}{n-p}}{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}} \qquad (8)$$

where the x_i are the individual elements (physical properties) contained within the test vector, $\bar{x}$ is the mean property value for all solutes or solvents, $\hat{x}_i$ represent the individual estimates obtained for the solute or solvent physical properties through the least-squares fit of the abstract factors to the test vector (i.e., the target test results), p is the number of abstract factors used in the target testing and n is the number of solvents or solutes in the test vector. We found by using this approach that we could compare similar scales. However, this approach still does not take into account the relative success in fitting scales with widely different inherent precision, or in fitting to an offset. When this normalized standard error is large, we cannot determine if the proposed term is important in describing the variance in the data set. In fact, if a test vector is highly correlated with a constant value (an offset), we would expect the NSE be fairly large. This is not to say that such terms do not contribute to

TABLE 5

Normalized standard errors (%) for alkane model

Test vector	Noise	
	5%	20%
Solute		
V_2	5.5	24
$V_2\delta_2$	4.2	18
$V_2\delta_2^2/RT + 0.3 \ln V_2$	2.5	13
$\mathbf{1}_2$ (offset)	0.071 [a]	0.26 [a]
Solvent		
δ_1^2	3.4	13
δ_1	6.7	25
1_1 (offset)	0.014 [a]	0.051 [a]
$\ln V_1$	34	130

[a] Standard error.

the variations in the data, but it is more difficult to determine their validity by target testing.

When PCA was performed on the data matrix with no noise, all 8 test vectors in Table 1 showed an excellent quality of fit with the abstract factors, as evidenced by very small NSE values. However, when noise was added to the data, different test vectors showed different fit qualities when fit to the abstract factors. This is illustrated in Table 5 where all of the NSE values are listed for the two synthetic alkane matrices with added noise contributions. This indicates that target testing of certain terms of Eqn. 4 were more susceptible to the presence of noise than others. For example, among the solute terms, the V_2 and $V_2\delta_2$ terms give a relatively poor fit quality, while for the solvent terms, the $\ln V_1$ term is very sensitive to the presence of noise. The specific reasons for these discrepancies in fit quality are not clear.

To determine if the fits for the 8 test vectors were significantly better than fits to a totally random vector, six additional random vectors were generated for each parameter (V_1, V_2, δ_1 and δ_2) over the same ranges that were used in the original simulation. These random vectors were used to construct six alternate test vectors for each linear term listed in Table 1, excluding the offsets. These alternative test vectors were fit to the abstract factors obtained from analysis of the synthetic data that contained 5 and 20% noise. The average normalized standard error for the 95% confidence interval was calculated for each of these test vectors. This information is shown in Table 6. The 95% confidence interval indicates the quality of fit that would be obtained if a real test factor could not be distinguished from a totally random parameter. From the six factors evaluated in this manner, the $\ln V_1$ term has a normalized standard error that is larger than the lower limit of the 95% confidence interval. Using this criterion, the $\ln V_1$ term is indistinguishable from an inappropriate test vector for the data matrix containing 20% noise, and the corresponding term for the solute, the offset $\mathbf{1}_2$, is also probably indistinguishable from an inappropriate test vector. Note that the SPOIL function indicated that the $\ln V_1$ term was a valid test vector, even though the random vector tests indicated that $\ln V_1$ was not distinguishable from a totally random vector.

Using the MOSCED equation (Eqn. 6) to simulate the logarithms of limiting activity coeffi-

TABLE 6

Normalized standard errors (%) for random vector target tests

Factor	Mean (5% noise)	Confidence interval [a] (5% noise)	Mean (20% noise)	Confidence interval [a] (20% noise)
V_2	103	76–130	128	87–169
$V_2\delta_2$	91	43–139	109	67–151
$V_2\delta_2^2/RT - 0.3 \ln V_2$	90	48–132	98	65–131
δ_1^2	108	88–128	102	81–123
δ_1	104	83–125	103	72–134
$\ln V_1$	97	85–109	113	84–142

[a] 95% confidence interval for six random vectors.

cients, three matrices were constructed containing 0, 5 and 20% normally distributed noise, as was done for the alkane model described above. These three matrices were analyzed with PCA and the abstract factors saved for target testing. The terms in Table 2 were used as test vectors. After eigenanalysis of the three MOSCED matrices with 0, 5 and 20% noise, the IND function, RSD and the *F*-testa consistently indicated 9, 7 and 6 significant factors, respectively, for the different noise levels. The NSE values for the test vectors are listed in Table 7.

After inspection of the NSE values for the MOSCED data, it appeared that any term that included a q^2 term gave a poor fit when compared to the other test vectors from the same matrix. This indicates that development of an appropriate dipolarity scale within this particular model framework may be very difficult, since the τ and q parameters always occur as a product. In addition, the test vector $\ln V_1$ has a relatively poor fit; this would indicate that identifying an appropriate V_1 scale may also be difficult.

TABLE 7

Normalized standard errors (%) for MOSCED model

Test vector	Noise	
	5%	20%
Solute		
V_2	5.2	17
$V_2\lambda_2$	5.5	17
$V_2q_2^2$	17	29
$V_2q_2^2\tau_2$	6.7	14
$V_2q_2^2\tau_2^2$	18	20
$V_2\alpha_2$	1.8	6.0
$V_2\beta_2$	2.1	6.3
$V_2(\lambda_2^2+\alpha_2\beta_2)/RT+0.3\ln V_2$	6.4	17
$\mathbf{1}_2$ (offset)	0.11 [a]	0.24 [a]
Solvent		
$\lambda_1^2+\alpha_1\beta_1$	8.5	16
λ_1	6.8	27
$q_1^2\tau_1^2$	26	27
$q_1^2\tau_1$	12	15
q_1^2	100	110
β_1	1.2	5.7
α_1	2.0	4.8
$\mathbf{1}_1$ (offset)	0.0064 [a]	0.026 [a]
$\ln V_1$	44	75

[a] Standard error.

TABLE 8

Normalized standard errors (%) for modified MOSCED model

Test vector	Noise	
	5%	20%
Solute		
V_2	3.5	16
$V_2(2\lambda_2+\tau_2)$	3.1	13
$V_2(\lambda_2^2+\lambda_2\tau_2+\tau_2^2+\alpha_2\beta_2)/RT$ $+0.3\ln V_2$	3.3	11
$V_2(\lambda_2+2\tau_2)$	2.7	12
$V_2\beta_2$	1.7	6.0
$V_2\alpha_2$	1.3	4.8
$\mathbf{1}_2$ (offset)	0.086 [a]	0.23 [a]
$V_2\lambda_2$	4.0	15
$V_2\tau_2$	3.7	14
$V_2(\lambda_2+\tau_2)$	2.7	12
Solvent		
$\lambda_1^2+\lambda_1\tau_1+\tau_1^2+\alpha_1\beta_1$	3.1	10
λ_1	11	36
$\mathbf{1}_1$ (offset)	0.010 [a]	0.031 [a]
τ_1	3.1	11
α_1	2.4	7.0
β_1	1.6	5.9
$\ln V_1$	24	86
$2\lambda_1+\tau_1$	7.3	23
$\lambda_1+2\tau_1$	3.7	12
$\lambda_1+\tau_1$	5.1	16

[a] Standard error.

The modified MOSCED model given by Eqn. 7 in the Theory section was also evaluated using 0, 5 and 20% noise added to the simulated activity coefficient values. The IND function, RSD and the *F*-test all indicated the presence of 7, 7 and 6 significant factors respectively. The NSE values for the test vectors are listed in Table 8. In this case, the NSE values are much more consistent for these test vectors, indicating that it should be much easier to identify meaningful scales for the solvent and solute dipolarity for this type of model.

Real data

PCA was used to analyze the 29×31 matrix containing 899 $\ln\gamma^\infty$ values described in the Experimental section. Approximately 25% of these values were originally estimated by UNIFAC or MOSCED; these values were improved using the factor analysis approach for compensating for

missing values, as described in the Theory section. During this process, all ln γ^{∞} values estimated from the factor analysis were examined and compared to the original experimental values contained in the database. Residuals which showed deviations greater than 20% were flagged; this allowed us to identify several typographical errors and suspicious values in the database. These errors were corrected before proceeding with the data analysis. The matrix returned by the missing data routine was analyzed by PCA and the abstract factors saved for target testing and for the cluster diagrams.

The results of principal component analysis of this data set show that the first 4 principal components describe 99.6% of the variation in the data. The first principal component describes 81.0% of the variance in the data set. The second, third and fourth principal components describe 15.7, 1.9 and 1.1% of the variance, respectively. The residual standard deviation should estimate the experimental uncertainty in the data set, if the correct number of factors has been chosen. PCA performed on the data set during the first iteration of the missing data routine (assuming four principal components) yielded an RSD of 0.188 ln units, or approximately a 20% error in the γ^{∞} values. This is consistent with the known error characteristics of the UNIFAC and MOSCED models. After the final iteration, the RSD was 0.102, reflecting an 11% error in the γ^{∞} values, which is more consistent with the uncertainties estimated for the experimental data contained in the matrix [19].

The correlations between each abstract factor and the ln γ^{∞} values for each solute were determined. Figures 2–5 show these correlations. Correlation coefficients in the range of 0.5–0.7 indicate a moderate correlation and correlation coefficients in the range 0.7–1.0 indicate a strong correlation.

Figure 2 shows the correlations for the first abstract factor with the solute ln γ^{∞} values. The correlation coefficients on this figure fall into 2 groups; those with correlation coefficients above 0.8 and those with correlation coefficients below 0.4. This indicates that the pattern in the most important principal component follows the behavior of the non-polar aliphatic, aromatic and halogenated aliphatic solutes. A majority of the solutes in this data set are non-polar and the first principal component reflects this bias. This is due to the fact that the first principal component always describes the greatest amount of variance in the data set. Figure 3 shows the correlations for the second abstract factor with the solute ln γ^{∞} values. The solutes that give a strong correlation with the ln γ^{∞} values are 1,4-dioxane, 2-butanone, acetone, acetonitrile, nitromethane and ethanol. These solutes are polar and are relatively strong hydrogen bonding bases. Figure 4 shows the correlations for the third abstract factor with the solute ln γ^{∞} values. This figure indicates a moderate correlation with methylene chloride and chloroform; however, benzene and toluene also show moderate correlation coefficients. This result could be due the polarizability of the solute contributing to this abstract factor.

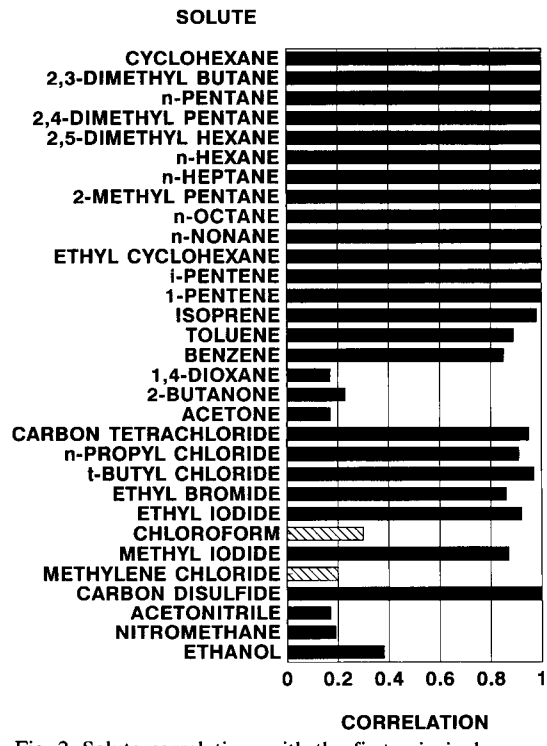


Fig. 2. Solute correlations with the first principal component (v_1). Solid bars represent positive correlation and hatched bars represent a negative correlation.

Figure 5 shows the correlations for the fourth abstract factor with the solute ln γ^{∞} values. This abstract factor is moderately correlated with chloroform and methylene chloride, while showing very little correlation with any of the other solutes. This suggests that chloroform and methylene chloride are unique in their behavior as compared to the other solutes in the data set. This could be because chloroform and methylene chloride are slightly acidic as hydrogen bond donors, but are only very weak hydrogen bond donors.

Figures 6 and 7 show the correlations with the first two abstract factors and the ln γ^{∞} values for each solvent. Figure 6 shows the results for the correlation of the first principal component with the solvent ln γ^{∞} values. This principal component shows a strong correlation with the polar compounds, while it shows a weak correlation with the non-polar solvents. Since there are more polar solvents present in the data matrix, this result is reasonable. Figure 7 shows the correlation of the second solute principal component with the solvent ln γ^{∞} values. This principal component has a strong correlation with the non-polar solvents while it shows a weak correlation with the polar compounds. The correlations for the third and fourth principal components are not given, as they showed essentially no correlation with any of the solvents.

Another means for presenting these results is to plot each of the solutes (or solvents) in the coordinate system generated by the principal components analysis. Figure 8 is a cluster diagram of the solutes plotted as a function of the first and second principal components. The non-polar solutes lie in a line parallel to the first principal component. The chlorinated aliphatic

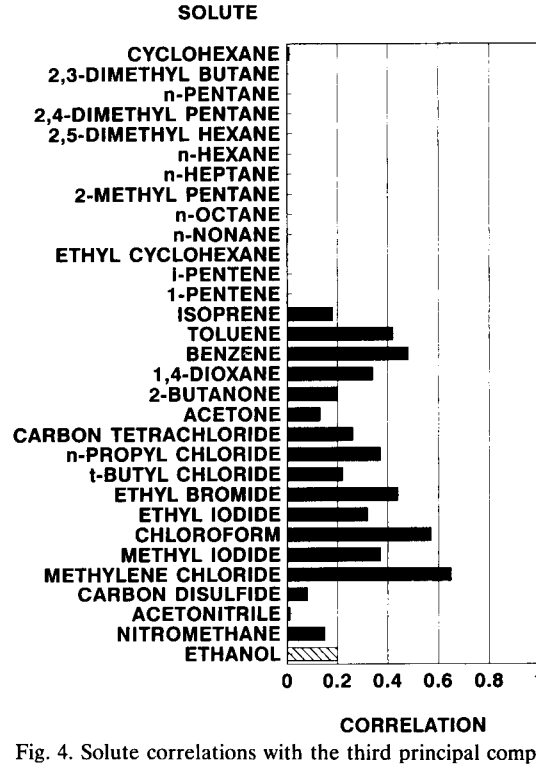


Fig. 4. Solute correlations with the third principal component (v_3). (see Fig. 2 for explanation).

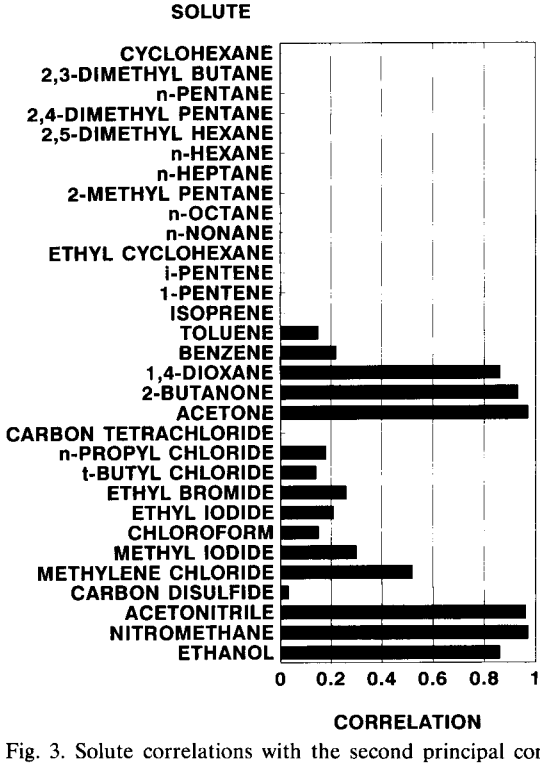


Fig. 3. Solute correlations with the second principal component (v_2). (see Fig. 2 for explanation).

compounds lie in a cluster centered around 0.1 on the first principal component axis. Chloroform and methylene chloride lie to the left of the chlorinated group. The two aromatic compounds included as solutes lie in the center of the chlorinated alkane cluster. The oxygen containing bases and the polar compounds form groups of their own. All of these groupings are consistent with the known chemistry of these compounds.

The principal components were used for target testing to evaluate the validity of the MOSCED model, and to explore the use of various pure component parameters within this model format. Since the MOSCED model expressed by Eqn. 6 yielded target test terms (shown in Table 2) that either confounded the q and τ terms with one another, or generated additive terms where the relative weighing of the terms would be difficult to ascertain for real physical parameters (i.e., $\lambda^2 + \alpha\beta$), we chose to evaluate the modified MOSCED model expressed by Eqn. 7. The formats for the target test terms for this model are given in Table 3.

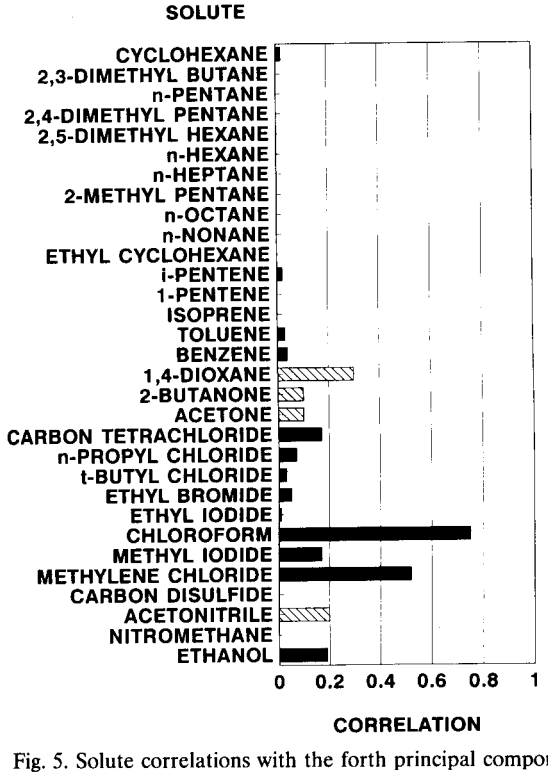


Fig. 5. Solute correlations with the forth principal component (v_4). (see Fig. 2 for explanation).

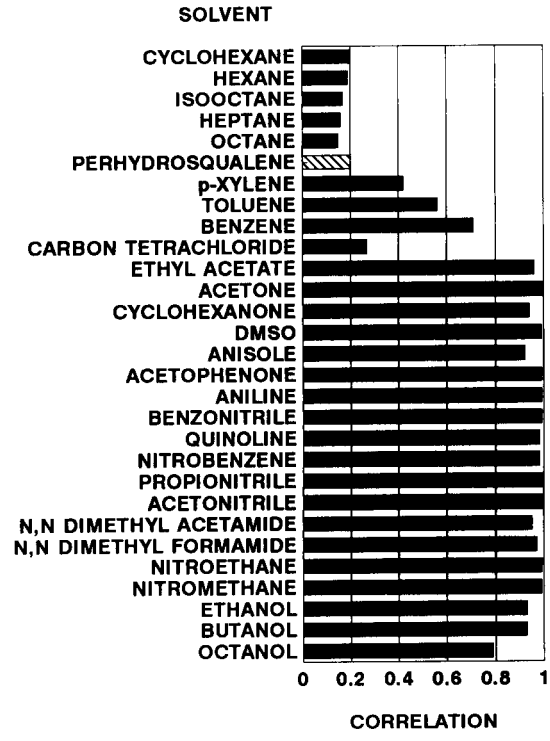


Fig. 6. Solvent correlations with the first principal component (u_1). (see Fig. 2 for explanation).

There are several experimental scales that might be used to represent the parameters that appear in this model. For the volume scale, the molar volume (V) and the hard sphere volume (V_B), as estimated according to Bondi [20], were tested. For representing the polarizability, the refractive index (n_D) or Onsager's function of the refractive index [$f(n_D)$] were examined [21]. For the polarity parameter, the solvatochromic π^*_{KT} scale, as developed by Kamlet et al. [22], which characterizes the dipolarity and polarizability of a solvent, was examined. An alternate π^* scale has been developed by Li et al. (π_2^{*C}) [23] which describes the dipolarity–polarizability of solutes. Hydrogen bond acidity and basicity of the sol-

vents were modeled by the α_{KT} and β_{KT} scales of Taft et al. [24]. Hydrogen bond acidity and basicity of the solutes were modeled by the α_2^h and β_2^h scales of Abraham et al. [25], or by the chromatographically-based α_2^C and β_2^C scales of Li et al. [23,26].

The results of the target testing are listed in Tables 9 and 10. A normalized standard error (NSE) of 40% or less was considered a good fit. NSE values between 40% and 80% appeared to be correlated with the test vectors when the fit was examined visually, and NSE values over 80% had a poor quality of fit. Test parameters with a small range compared to their offsets tend to have higher NSEs, however, a high NSE does not necessarily imply that the factor does not contribute to the overall model but that it is more difficult to identify using target testing. The NSE values obtained were compared to NSE values

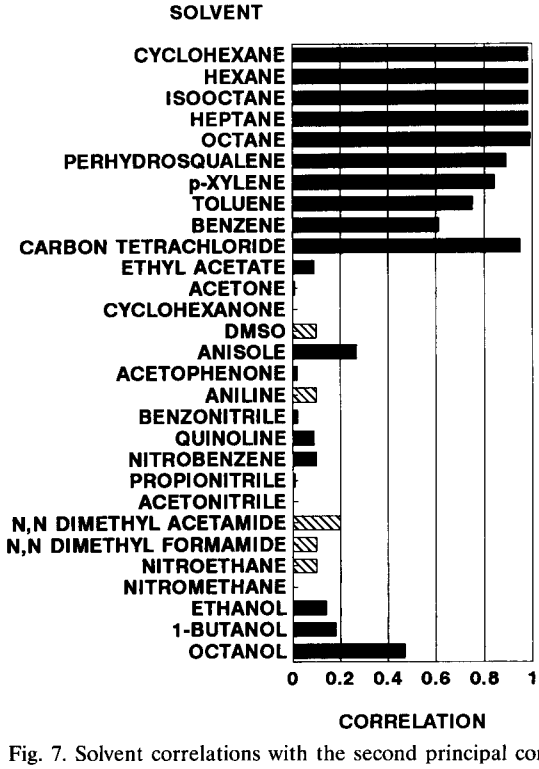


Fig. 7. Solvent correlations with the second principal component ($\mathbf{u}_2$). (see Fig. 2 for explanation).

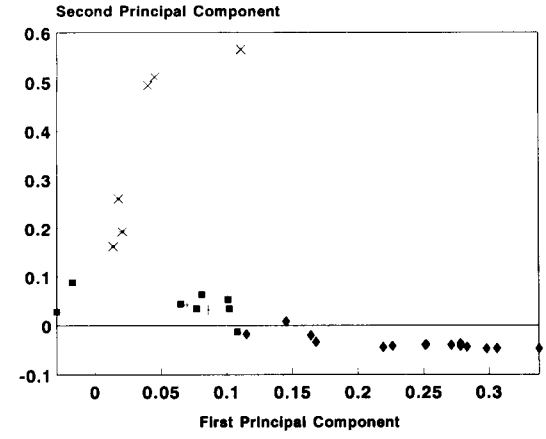


Fig. 8. Cluster diagram of the solutes plotted as a function of the first and second principal components. (♦) Non-polar compounds (cyclohexane, 2,3-dimethylbutane, *n*-pentane, 2,4-dimethylpentane, 2,5-dimethylhexane, *n*-hexane, *n*-heptane, 2-methylpentane, *n*-octane, *n*-nonane, ethylcyclohexane, isopentene, 1-pentene, isoprene and carbon disulfide). (+) Aromatic compounds (toluene and benzene). (*) Oxygen containing base compounds (1,4-dioxane, 2-butanone and acetone). (■) Halogenated compounds (carbon tetrachloride, *n*-propyl chloride, *tert*-butyl chloride, ethyl bromide, ethyl iodide, chloroform, methyl iodide and methylene chloride). (×) Polar compounds (acetonitrile, nitromethane and ethanol).

obtained from fits to random test vectors with the same offset and standard deviation as the test vector. The results listed are the 95% confidence intervals for the NSE values for 6 random test vectors. These results indicate that while many of the NSE values are large, the test vectors can account for more of the variation in the data than a purely random vector. The standard errors of the test vectors are also given in Table 9. Another approach for evaluating the success of target testing is to compare the standard error of fit to the experimental error in the target test vector parameters.

In general, the factors for the solutes fit slightly better than the corresponding terms for the solvents. This may be related to the observation that the third and fourth solvent factors do not yield significant correlations with any of the individual solvents that were studied. There does not ap-

TABLE 9

Target tests for modified MOSCED model: solutes [a]

Parameter	NSE (%)	SE [b]	Range [c]	95% C.I. [d] (%)
	V_2			
V_B	60	12	28–99	86–164
V	60	21	52–180	86–158
	$V_2\lambda_2$			
$Vf(n_D)$	60	4.0	9.3–35	118–138
	$V_2(2\lambda_2+\tau_2)$, $V_2(\lambda_2+2\tau_2)$, $V_2\tau_2$			
$V\pi_2^{*C}$	39	8.2	−29–36	98–110
$V_B\pi_2^{*C}$	39	4.4	−16–20	97–105
π_2^{*C}	21	0.050	−0.18–0.67	86–110
$V^{1/2}\pi_2^{*C}$	30	0.65	−2.1–4.9	93–111
$V\pi_{KT}^{*}$	51	13	−10–63	99–109
$V_B\pi_{KT}^{*}$	51	6.8	−5.5–33	95–105
π_{KT}^{*}	37	0.12	−0.08–0.85	96–108
	$V_2(\lambda_2^2+\lambda_2\tau_2+\alpha_2\beta_2)+0.3\ln V_2$			
$V\delta_H^2/RT+0.3\ln V$	109	2.6	9.4–20	124–196
$V_B\delta_H^2/RT+0.3\ln V_B$	101	1.5	5.3–12	129–221
$V\delta_H^2$	101	1400	4800–11 000	122–214
δ_H^2	48	15	45–170	100–124
$V^{1/2}\delta_H^2$	75	140	460–1200	105–179
	$V_2\beta_2$			
$V\beta_{KT}$	81	9.3	0–43	96–114
$V_B\beta_{KT}$	81	5.1	0–24	90–112
β_{KT}	64	0.099	0–0.48	94–110
$V^{1/2}\beta_{KT}$	73	0.97	0–4.5	90–116
$V\beta_2^h$	78	9.0	0–43	85–117
$V_B\beta_2^h$	77	4.8	0–24	96–112
β_2^h	61	0.096	0–0.50	88–116
$V^{1/2}\beta_2^h$	70	0.93	0–4.5	93–115
$V\beta_2^C$	87	13	0–67	93–107
$V_B\beta_2^C$	88	7.5	0–38	96–110
β_2^C	77	0.15	0–0.79	85–113
$V^{1/2}\beta_2^C$	80	1.5	0–7.5	81–111
	$V_2\alpha_2$			
$V\alpha_{KT}$	34	3.7	0–48	90–108
$V_B\alpha_{KT}$	33	1.9	0–25	80–112
α_{KT}	25	0.044	0–0.82	83–113
$V\alpha_2^h$	36	1.7	0–19	90–110
$V_B\alpha_2^h$	35	0.87	0–10	80–116
α_2^h	28	0.021	0–0.33	86–114
$V\alpha_2^C$	34	1.4	0–19	95–109
$V_B\alpha_2^C$	33	0.70	0–10	83–113
α_2^C	26	0.017	0–0.33	84–114
$V^{1/2}\alpha_2^C$	27	0.15	0–2.5	85–111
	$\mathbf{1}_2$			
Offset	–	0.23	–	–

[a] See List of symbols. [b] Standard error. [c] Minimum and maximum parameter values for the compounds in the data set. [d] 95% confidence interval of the NSE values for 6 random vectors.

TABLE 10

Target tests for modified MOSCED model: solvents [a]

Parameter	NSE (%)	SE [b]	Range [c]	95% C.I. [d] (%)
	$(\lambda_1^2 + \lambda_1\tau_1 + \tau_1^2 + \alpha_1\beta_1)$			
δ_H^2	74	28	47–190	88–152
δ_H^2/V	50	0.38	0.13–3.0	77–127
$\delta_H^2/V^{1/2}$	54	2.9	2.9–22	83–125
	λ_1			
n_D	430	0.31	1.35–1.63	282–572
$f(n_D)$	220	0.051	0.18–0.26	167–271
	$(2\lambda_1 + \tau_1)$, $(\lambda_1 + 2\tau_1)$, τ_1			
π_2^{*C}	47	0.16	−0.16–1.0	79–133
π_2^{*C}/V	29	0.0013	−0.001–0.014	89–123
$\pi_2^{*C}/V^{1/2}$	36	0.013	−0.01–0.12	92–126
π_{KT}^*	55	0.19	−0.04–1.0	84–126
π_{KT}^*/V	36	0.0017	−0.003–0.016	84–128
$\pi_{KT}^*/V^{1/2}$	42	0.016	−0.004–0.12	97–115
	$\mathbf{1}_1$			
Offset	–	0.20	–	–
	α_1			
α_2^h	72	0.080	0–0.33	94–110
α_2^h/V	65	0.0009	0–0.0057	84–114
$\alpha_2^h/V^{1/2}$	67	0.0080	0–0.043	77–125
α_2^C	73	0.078	0–0.33	88–108
α_2^C/V	68	0.0009	0–0.0057	76–116
$\alpha_2^C/V^{1/2}$	70	0.0079	0–0.043	85–115
α_{KT}	72	0.18	0–0.82	97–111
α_{KT}/V	66	0.0022	0–0.014	88–116
$\alpha_{KT}/V^{1/2}$	68	0.019	0–0.11	92–108
	β_1			
β_2^h	50	0.12	0–0.78	81–125
β_2^h/V	35	0.0011	0–0.011	93–115
$\beta_2^h/V^{1/2}$	39	0.010	0–0.093	84–126
β_2^C	37	0.13	0–1.42	85–117
β_2^C/V	31	0.0014	0–0.020	77–123
$\beta_2^C/V^{1/2}$	31	0.013	0–0.17	92–116
β_{KT}	64	0.19	0–0.76	83–119
β_{KT}/V	42	0.0015	0–0.011	99–119
$\beta_{KT}/V^{1/2}$	50	0.016	0–0.090	90–114
	$\ln V_1$			
$\ln V_B$	200	0.92	3.3–5.7	159–255
$\ln V$	230	1.03	4.0–6.3	193–285

[a–d] See footnotes to Table 9.

pear to be a significant difference in the quality of fits when using molar volume or the Bondi volume scale. This should be expected since these scales are very collinear. The parameters used for modeling the polarizability of the solvent gave particularly poor NSE values, but this is largely due to the limited range of these parameters. In the case of the solutes, the $V\pi^*$ terms did not fit

as well as π^*, and for the solvents, the π^* terms did not fit as well as π^*/V, indicating that the π^* scales may require normalization by the volume or the square root of the volume, as might be suspected from a dimensional analysis. Terms for the hydrogen bonding acidity fit well for the solutes in the data set but did not fit as well for the solvents. Terms for the hydrogen bonding basicity fit well for the solvents in the data set but did not fit as well for the solutes. These results agree since the solute acidity term and the solvent basicity term appear in the same term, as seen in Table 3. In addition, for the solutes, the $V\alpha$ and $V\beta$ terms did not fit as well as α and β, and for the solvents, the α and β terms did not fit as well as α/V and β/V, indicating that the hydrogen bonding terms may also require normalization by the volume. In theory, the solute-based scales (C and h) should better represent solute properties, while the solvent-based scales (KT) should better represent the solvent properties, however, the NSE values indicate that the solute-based scales (C and h) fit as well as the solvent-based scales (KT) for the solvents.

Conclusions

This paper illustrates a novel use of factor analysis to characterize infinite dilution activity coefficients. Factor analysis aided in the detection of outliers and typographical errors in the database. In addition, by using factor analysis with synthetic data matrices generated to study the different models for infinite dilution activity coefficients, we were able to determine the number of factors for each model and identify factors which may be more susceptible to noise. This work with the synthetic data matrices also indicated which model for infinite dilution activity coefficients may be the most desirable to use to develop different scales for the physicochemical parameters (i.e., volume, dipolarity, dispersion and hydrogen bonding interactions).

Factor analysis and target testing provide a basis to study different model forms of the MOSCED equation and different physicochemical scales by using a database of real infinite dilution activity coefficients and physical properties. The quality of fit from the test parameters suggests that some scales may need to be normalized by the volume, as indicated by the lower normalized standard errors. The scope of the target testing described here was limited by the noise in the original data set. However, this paper demonstrates that factor analysis methods can be used successfully in the evaluation of very complex models.

SCR and RBP would like to acknowledge the National Science Foundation (Grant CHE-8921315) for supporting this research. MJH and CAE would like to acknowledge the financial support of DuPont Chemical Co., Amoco Chemical Co. and the National Science Foundation. PWC would like to acknowledge support from the National Science Foundation.

LIST OF SYMBOLS

Subscripts

1 Solvent
2 Solute

Symbols

V Volume or molar volume.
V_B Hard sphere volume according to Bondi [20].
δ Hildebrand's solubility parameter for synthetic alkane model.
δ_H Hildebrand's solubility parameter (experimentally determined).
n_D Refractive index.
$f(n_D)$ Onsager function of refractive index [21] $f(n_D) = (n_D^2 - 1)/(2n_D^2 + 1)$.
λ Dispersion parameter for synthetic MOSCED models.
q Induction parameter for synthetic MOSCED model.
τ Polar parameter for synthetic MOSCED models.
π_2^{*C} Combination dipolarity–polarizability parameter experimentally determined by Li et al. [23].
π_{KT}^* Combination dipolarity–polarizability parameter experimentally determined by Kamlet et al. [22].

α Hydrogen bonding acidity parameter for synthetic MOSCED models.

α_2^C Hydrogen bonding acidity parameter experimentally determined by Li et al. [23].

α_{KT} Hydrogen bonding acidity parameter experimentally determined by Taft et al. [24].

α_2^h Hydrogen bonding acidity parameter experimentally determined by Abraham et al. [25].

β Hydrogen bonding basicity parameter for synthetic MOSCED models.

β_2^C Hydrogen bonding basicity parameter experimentally determined by Li et al. [26].

β_{KT} Hydrogen bonding basicity parameter experimentally determined by Taft et al. [24].

β_2^h Hydrogen bonding basicity parameter experimentally determined by Abraham et al. [25].

1 Offset vector [1 1 1 ... 1 1].

Notation for linear algebra

Scalars are represented by lower case italics. Column vectors are represented by bold, lower case letters. Row vectors are represented by bold, lower case letters with the superscript T for transpose. Matrices are represented by bold, upper case letters.

REFERENCES

1 S. Zeck, Fluid Phase Equilib., 70 (1991) 125.
2 J.M. Prausnitz, R.N. Lichtenthaler and E.G. De Azevedo, Molecular Thermodynamics of Fluid-phase Equilibria, Prentice-Hall, Englewood Cliffs, NJ, 2nd edn., 1986.
3 M. Roth and J. Novak, J. Chromatogr., 258 (1983) 23.
4 L.B. Schreiber and C.A. Eckert, Ind. Eng. Chem. Process Des. Dev., 10 (1971) 572.
5 A. Hussam and P.W. Carr, Anal. Chem., 57 (1985) 793.
6 E.R. Thomas, B.A. Newman, G.L. Nicolaides and C.A. Eckert, J. Chem. Eng. Data, 27 (1982) 233.
7 J.M. Miller, Chromatography: Concepts and Contrasts, Wiley, New York, 1988, p. 150.
8 J.C. Bastos, M.E. Soares and A.G. Medina, Ind. Eng. Chem. Res., 27 (1988) 1269.
9 J.H. Park and P.W. Carr, Anal. Chem., 59 (1987) 2596.
10 E.R. Thomas and C.A. Eckert, Ind. Eng. Chem. Process Des. Dev., 23 (1984) 194.
11 W.J. Howell, A.M. Karachewski, K.M. Stephenson, C.A. Eckert, J.H. Park, P.W. Carr and S.C. Rutan, Fluid Phase Equilib., 52 (1989) 151.
12 A.P. Dempster, N.M. Laird and D.B. Rubin, J. Roy. Stat. Soc. B, 39 (1977) 1.
13 E.R. Malinowski, Anal. Chem., 49 (1977) 612.
14 E.R. Malinowski, Anal. Chem., 49 (1977) 606.
15 E.R. Malinowski, J. Chemom., 3 (1988) 49.
16 E.R. Malinowski, Factor Analysis In Chemistry, Wiley, New York, 2nd ed., 1991.
17 J.H. Hildebrand, J.M. Prausnitz and R.L. Scott, Regular and Related Solutions, Van Nostrand Reinhold, New York, 1970.
18 S.C. Rutan and P.W. Carr, (1988) unpublished results.
19 M.J. Hait, C.L. Liotta, C.A. Eckert, D.L. Bergmann, A.M. Karachewski, A.J. Dallas, D.I. Eikens, J.J. Li, P.W. Carr, R.B. Poe and S.C. Rutan, Ind. Eng. Chem. Res., submitted for publication.
20 A. Bondi, Physical Properties of Molecular Crystals, Liquids, and Glasses, Wiley, New York, 1968.
21 L. Onsager, J. Am. Chem. Soc., 58 (1936) 1486.
22 M.J. Kamlet, J.L. Abboud and R.W. Taft, J. Am. Chem. Soc., 99 (1977) 6027.
23 J. Li, Y. Zhang, A.J. Dallas and P.W. Carr, J. Chromatogr., 550 (1991) 101.
24 R.W. Taft, J.M. Abboud, M.J. Kamlet and M.H. Abraham, J. Sol. Chem., 14 (1985) 153.
25 M.H. Abraham, P.L. Grellier, D.V. Prior, P.P. Duce, J.J. Morris and P.J. Taylor, J. Chem. Soc., Perkin Trans. II, (1989) 699.
26 J. Li, Y. Zhang, H. Ouyang and P.W. Carr, J. Am. Chem. Soc., (1992) in press.

Analytica Chimica Acta, 277 (1993) 239–253
Elsevier Science Publishers B.V., Amsterdam

DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures

S. Wold, J. Jonsson, M. Sjöström, M. Sandberg and S. Rännar
Research Group for Chemometrics, Umeå University S-901 87 Umeå (Sweden)

(Received 3rd September 1992)

Abstract

Biopolymer sequences (e.g., DNA, RNA, proteins and polysaccharides) and chemical processes (e.g., a batch or continuous polymer synthesis run in a chemical plant) have close similarities from the modelling point of view. When a set of sequences or processes is characterized by multivariate data, a three-way data matrix is obtained. With sequences the position and with processes the time is one direction in this matrix. The multivariate modelling of this matrix by principal component analysis (PCA) or partial least-squares (PLS) methods for the following purposes is discussed: classification of sequences; quantitative relationships between sequence and biological activity or chemical properties; optimizing a sequence with respect to selected properties; process diagnostics; and quantitative relationships between process variables and product quality variables. To obtain good models, a number of problems have to be adequately dealt with: appropriate characterization of the sequence or process; experimental design (selecting sequences or process settings); transforming the three-way into a two-way matrix; and appropriate modelling and validation (modelling interactions, periodicities, "time series" structures and "neighbour effects"). A multivariate approach to sequence and process modelling using PCA and PLS projections to latent structures is discussed and illustrated with several sets of peptide and DNA promoter data.

Keywords: Process analysis/on-line analysis; Pattern recognition; Biopolymer sequences; DNA sequences; Multivariate modelling; Partial least squares; Peptide sequences; Principal component analysis

With the increasing analytical abilities of biochemists, analytical chemists, microbiologists and molecular biologists, the sequences of peptides, proteins, DNA and other biopolymers are being determined at an increasing rate. To bring some order to these data and to use them for purposes such as the understanding of the evolutionary relationships between organisms and relationships between sequence and biological activity and chemical properties, various kinds of models are needed. With the complexity of sequences, i.e., many positions times several properties, these models will necessarily be multivariate [1,2].

Correspondence to: S. Wold, Research group for Chemometrics, Umeå University, S-901 87 Umeå (Sweden).

The monitoring and control of chemical processes is another area that is currently undergoing a data explosion. Sensors and on-line instrumentation provide multitudes of data characterizing the state of the process at any given point in time [3]. A common objective is to relate these time sequences of data to qualities of the resulting products of the process, such as yield, purity, crystal size, polymer strength and viscosity. Again, multivariate modelling provides the tools for handling these data analysis problems [4–6].

Interestingly, these two application areas of multivariate modelling show close similarities. A biopolymer sequence, such as a peptide or a DNA, has a univariate direction from beginning to end, analogous to a time sequence of process

data. The consecutive observations of a time sequence of process data are usually not independent. This has necessitated the use of special types of models, time series models [7], for the analysis of process data. It may be of interest to apply these models also to polymer sequences to investigate whether nature introduces "time series patterns" in the biological sequences such as neighbour effects, periodicities and auto-correlations.

The modelling of a set of polymer sequences characterized by a multivariate description of each position leads to the analysis of a three-way matrix (sequence × position × descriptor). The same situation appears in the modelling of a set of, say, batch process time sequences, with multivariate observations at each time point (process × time × variable). Before the analysis with standard methods, these three-way matrices must be transformed into two-way matrices, which can be done in different ways. These ways are not equivalent, but correspond to different assumptions about the underlying action mechanisms of the modelled sequences or processes.

In this paper some aspects of sequence and process modelling are discussed, in particular their common features. Examples of peptide and nucleic acid sequence models are used as illustrations.

SCOPE OF SEQUENCE AND PROCESS MODELLING

The first problem in any investigation is to collect quantitative and relevant data characteriz-

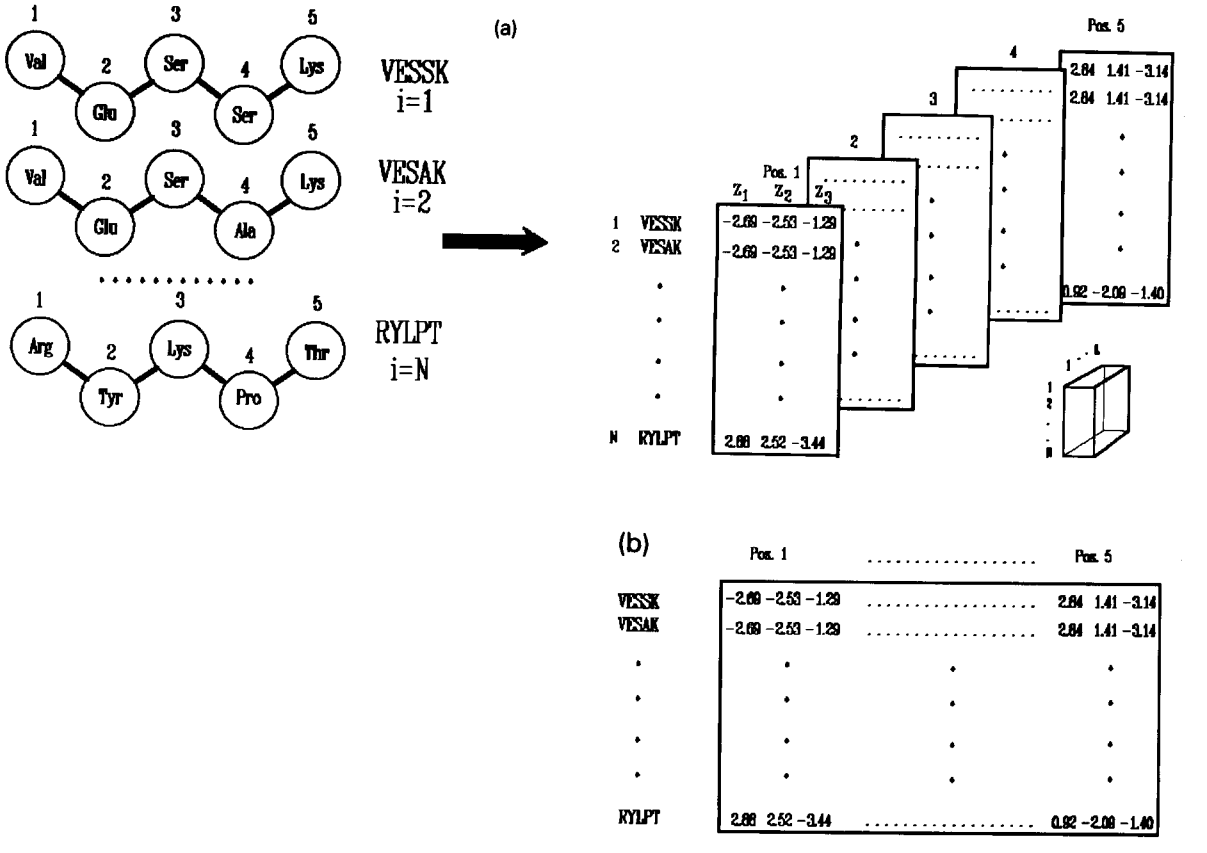


Fig. 1. (a) A set of biopolymer sequences can be quantified by using descriptor scales such as z_1, z_2 and z_3 for each varying position. The resulting descriptor data matrix (**X**) is three-way (sequence × position × descriptor). (b) In the case of sequences of the same length, the three-way table can be directly "unfolded" to give a two-way table with several columns for each sequence position.

ing the investigated objects (here sequences). This gives a data matrix in which each row corresponds to an object (sequence) and each column corresponds to a variable, calculated or measured on the objects. With sets of polymer or time sequences, this is often a three-way data matrix (Figs. 1 and 2), as will be discussed below.

Once quantitative and relevant data have been obtained, one needs to analyse the data, i.e., relate them to various more or less clearly formulated objectives. These objectives can be grouped into three broad categories: overview, classification and quantitative modelling.

Overview; finding groups, patterns

When little is known about the problem or system under investigation, the scope of the analysis is usually to obtain an overview of the objects in the data (here biopolymer or process time sequences). The result is a "map" of the similarities and dissimilarities of the objects, a model of the system. Sometimes groups of objects (classes) can be distinguished in this map, but often the objects are distributed over the map without any clear gaps. Other "patterns" may still be seen in the map, such as trends in time or other coordinates.

The analysis also informs on which variables (object characteristics) are related to directions in the map and other discernible patterns.

The data analysis tools useful for this type of analysis are scatter plots when the total number of variables, K, is smaller than ca. 5, and principal components analysis (PCA) and factor analysis (FA) when K is larger [8,9].

Diagnostics

The model obtained in PCA or FA can be used in a second stage of the data analysis to scrutinize new objects and their relationships to

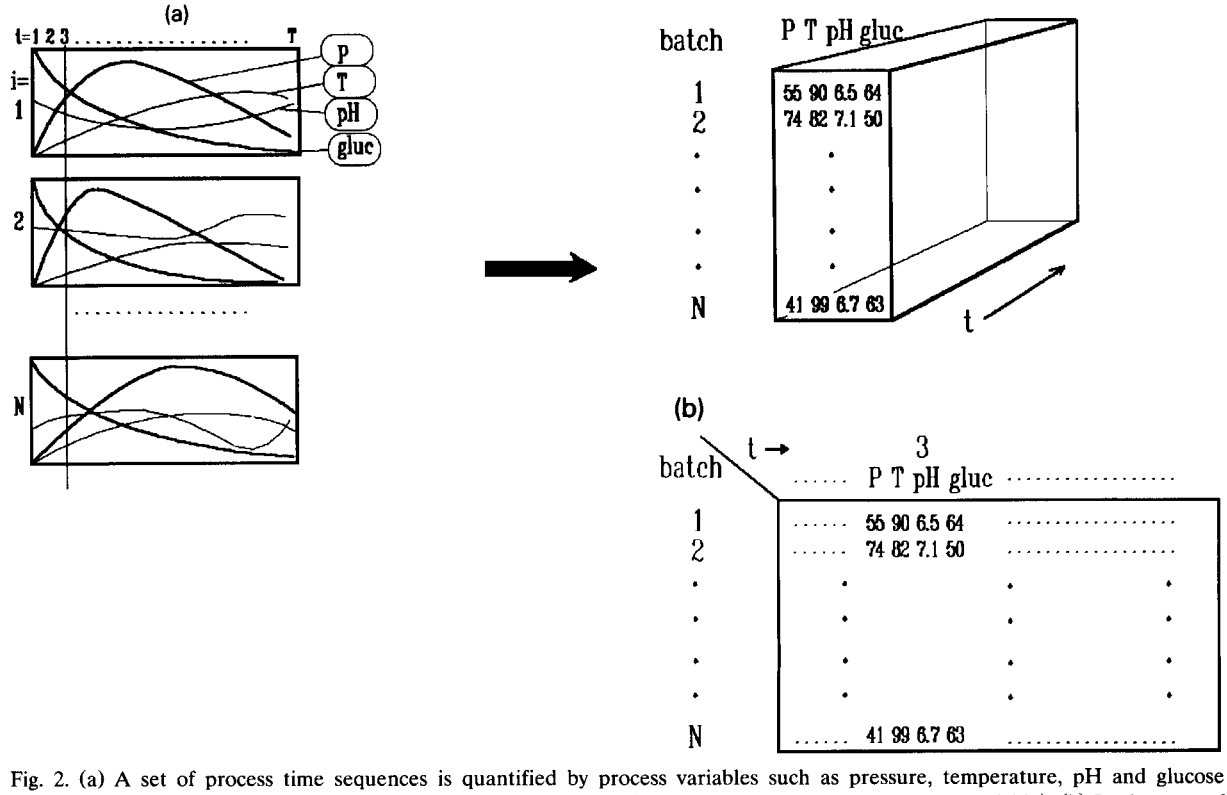


Fig. 2. (a) A set of process time sequences is quantified by process variables such as pressure, temperature, pH and glucose concentration. The resulting process data matrix (**X**) is three-way (process sequence × time × process variable). (b) In the case of process time sequences of the same length, the three-way table can be directly "unfolded" to give a two-way table with several columns for each process time point.

those previously analysed and to the model. This gives a "diagnosis" of each new object, whether it is similar to the set of previously modelled objects and, if so, in which region of the map the object falls.

For analytical chemistry, correlations between the position in the map and chromatographic retention, solubility, etc., are of interest.

With biopolymer sequences it may be of interest to correlate the inter-point distances in the map to evolutionary or genetic resemblance. Relationships between the place in the map and various diseases are relevant for medical diagnostic objectives.

Analogously, multivariate process diagnosis has recently started to develop, where patterns in multivariate time sequences are used to diagnose the state of the process, whether it is running normally or whether problems are occurring [4–6,10].

Often the diagnostic objective is formulated in terms of a classification, as discussed further below.

Classification into known classes (groups)

When more is known about the investigated system, this knowledge is often formulated as a class structure; peptides are classified according to their type of biological activity as enkephalins, bradykinins, signal peptides, etc. A class may be subdivided into sub-classes, e.g., signal peptides are divided into periplasmic or outer membrane and other types.

A process may be divided in classes such as "normal and under control", "raw material problems", "temperature problems in reactor 1" and "stack-up problems before reactor 2".

The scope of the analysis now is usually to see how the data collected for the objects relate to this class structure, and thereafter to classify new objects accordingly. The result of the analysis is still a "map" of the similarities and dissimilarities of the objects, a model of the system, often with sub-maps for the different classes. Again, the analysis will inform on which variables are important for the class separation and which are not. Also, the model informs about the degree of separation of the classes, which are well resolved and not, etc.

The model can then be used in a second stage of the data analysis to classify new objects. The result may be one of three: that an object is uniquely assigned to one of the classes on a certain probability level, or that the object is similar to several classes or that it is similar to none of the classes, an outlier, a new type.

The tools for classification, or pattern recognition, or discriminant analysis as this analysis is often called, are many, and are covered in reviews and textbooks [11,12]. When the number of characterizing variables (K) is larger than the number of objects (N), which is common in the analysis of sequences, projection methods such as SIMCA [13] and PLS discriminant analysis [14,15] are advantageous.

Quantitative relationships between sequence and biological activity or process variables and product properties

On the highest ambition level of data analysis, the scope is to find relationships between the characteristics of the objects (sequences) and other quantitative variables (response variables), e.g., the measured levels of biological activity variables, enzyme binding constants or other physico-chemical variables, such as chromatographic retention times and solubility. With processes, the responses are "properties" of the product and process such as yield, purity, stability, strength and viscosity, including "negative" results such as costs and pollution levels.

Often the data analysis is a combination of a classification and a quantitative modelling: the first objective is to find the class or type of an object, and thereafter to obtain a prediction of values of quantitative property variables that are defined for this class.

In the analysis, a quantitative model relating the characteristics of the objects to the response variables is derived with the help of a training set of objects with known values of the response variables. If several classes are at hand, one usually develops one such model for each class.

The developed model(s) are then used to predict the response values for new objects, after a classification if so needed.

Projection methods such as PLS (partial least-squares projection to latent structures) are suitable tools for this type of analysis [1,2,4,11,14–17]. In simple cases when few variables are needed to characterize the objects, and when statistical experimental design [18,19] has been applied for the selection of the training set, and when the number of responses is small (say a maximum of 5) and independent, multiple regression can be used for the model development and data analysis [20].

Optimization

A common objective is to achieve optimum properties of a biopolymer sequence or optimum quality and yield of a process. The models can be used for this purpose in several ways. With simple models, one can search mathematically for their maxima and minima, but with more complicated models simulations are necessary.

SELECTING A SET OF SEQUENCES (DESIGN OF A TRAINING SET)

To develop a model of a family of sequences, it is necessary to have a basis in the form of a representative sub-sample of family members. The selection of this sub-sample is far from trivial, and only recently has a rational approach become available with the methodology of statistical experimental design [2,18,19,21]. The intuitive way of generating a sub-set by modifying one position at a time in a "lead sequence" does not give a sub-set with sufficient information for good model development [2,21]. Unfortunately, this intuitive approach is still used in most applications, which is a major reason for the poor quality of structure–activity models in general.

EXAMPLES

To illustrate the principles of biopolymer sequence modelling, some sets of short peptides and a set of fairly short DNA sequences that were previously modelled by position based quantitative structure–activity relationships (QSARs) will be used. These sets of biopolymers will be modelled in two ways, with position-based description (where possible), and auto- and cross-covariance (ACC, see below)-based description which is alignment independent. In addition, some simulations have been made to demonstrate the non-equivalence between these two ways of modelling.

TABLE 1

Results of the PLS modelling of (1) the position-based (pos.) and (2) the auto-correlation (ACC)-transformed biopolymer sequence data [a]

Parameter	Ex. 1, enkephalins	Ex. 2, bradykinins	Ex. 3, signal peptides	Ex. 4, DNA promoters	Ex. 5, simulations
Length (L)	5	9	18 –28	68	5, 10
No. of sequences (N)	31	43	22	25	16, 32
No. of X-variables, K (pos.)	16	34	60	195	15, 30
R^2, mult. corr. (pos.)	0.59	0.52	0.65	0.85	0.88, 0.85
Q^2, R.S.D.–mult. corr. (pos.)	0.46	0.25	0.12	0.40	0.70, 0.30
No. of X-variables, K (ACC)	48	48	54	27	27, 27
R^2, mult. corr. (ACC)	0.58	0.48	0.91	0.79	0.30, 0.31
Q^2, R.S.D.–mult. corr. (ACC)	0.49	0.29	0.63	0.41	0.0, 0.0

[a] R^2 and Q^2 refer to the ordinary and cross-validated multiple correlation coefficients. $R^2 = 1 - [SS\ Y\text{-resid}/SS\ Y]$, where SS denotes sum of squares, Y-resid the residuals and Y the observed values of the response variable. The number of X-variables in the models is denoted by K and the number of cases by N.

(1) MVD enkephalins (length L = 5)

These $N = 31$ pentapeptides are of interest because they contain some D-amino acids in the sequences of mainly L-amino acids. They were previously modelled by position-based QSARs [22], where each of the four varying positions (position 1 is constant) were coded by the three z-scales of Hellberg [23] and Jonsson et al. [24]. An additional indicator variable was included to code for the D- or L-form of the amino acid. Hence the position-based description takes sixteen X-variables: four positions × (three z-scales and one indicator). A PLS model with two significant components explains $R^2 = 0.59$ of the Y-variation (enkephalin activity), with a cross-validated multiple correlation coefficient (Q^2) of 0.46 (see Table 1).

(2) Bradykinins (L = 9)

These $N = 43$ nonapeptides (length $L = 9$) were previously modelled [22] in the same way as the enkephalins in example 1. These bradykinins also had both D- and L-amino acids, except in positions 5 and 8, which had only L-forms. Hence, a position-based characterization with the same four variables as in the example 1 takes 34 variables [$(9 \times 3) + (7 \times 1)$]. A two-dimensional PLS model gives $R^2 = 0.52$ and $Q^2 = 0.25$.

(3) Signal peptides ($18 \leqslant L \leqslant 28$)

Signal peptides are N-terminal peptides sitting on proteins that, after their synthesis in the cytoplasmic domain of the cell, are translocated to a membrane or through one or several membranes. These signal peptides usually consist of 15–30 amino acid residues. Previously a set of *Escherichia coli* signal peptides were analysed and it was shown that the sequences do indeed contain patterns related to the site to which they direct the protein [25]. This analysis was made with a modified position-based description of the sequences based of the three z-scales [23,24].

The multi-positional description of peptide sequences used in QSAR studies [2,22,23] is not applicable here because the numbers of amino acids in the sequences are different. Hence the

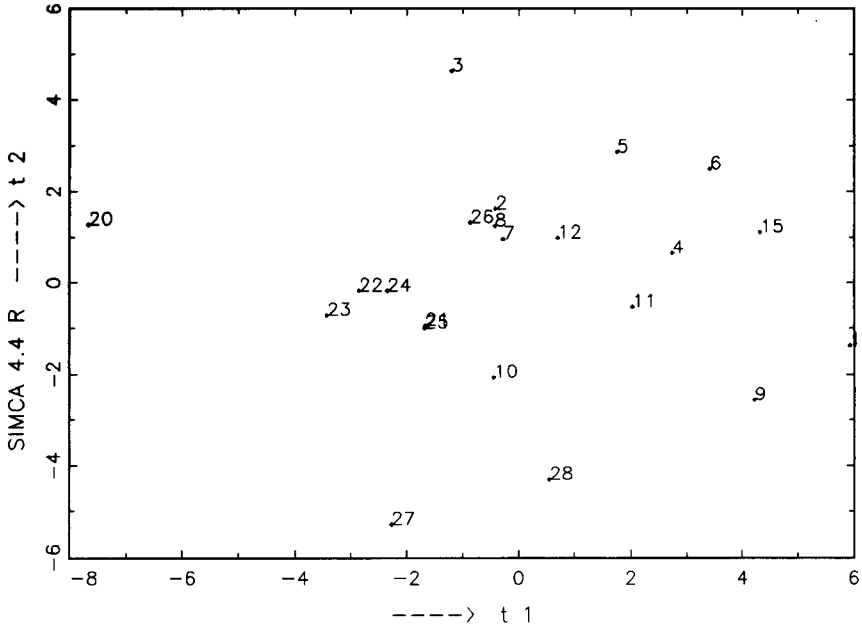


Fig. 3. Score plot of the PLS discriminant model of the signal peptides quantified by the "field" approach [25]. Nos. 1–15 are periplasmic space signal peptides and Nos. 20–28 are outer membrane signal peptides. The periplasmic peptide 10 and outer membrane peptide 26 are seen to fall in wrong regions.

sequences would be represented by different numbers of variables and it is not obvious which variables correspond to each other. In an effort to circumvent these problems, the so-called field approach was previously developed, where the two-dimensional matrix of one sequence is projected down on a vector with specified length. Hence the result is a specific number of variables for each sequence, independent of sequence length. The field approach is described in detail in a previous paper [25].

The field approach was used to generate a data matrix for 43 signal peptide sequences with a length between 18 and 32 amino acids. The final non-cytoplasmic locations of the corresponding proteins in *E. coli* were used as class labels. The number of classes was five: periplasmic space, outer membrane, pili, toxins and inner membrane proteins. The difference between pair-wise classes was investigated with PLS discriminant analysis. Significant differences according to cross-validation were obtained between most pairs of classes. However, the cross-validated explained variance of the discriminant variable was fairly low, which made the classification of sequences with unknown class membership uncertain.

Here the results from a PLS discriminant analysis based on the field approach are compared with those based on the auto- and cross-covariances of the *z*-scales coding the amino acid positions. To limit this preliminary study, only two classes of signal peptide sequences are considered, namely those of periplasmic space proteins and those of outer membrane proteins. The sequences are Nos. 1–22 in Table I in [25]. The location of the protein corresponding to protein No. 20 (MolA) has been reclassified from an outer membrane to a periplasmic protein since the previous analysis.

A PLS discriminant analysis [14,15] was done on a data matrix with 60 variables generated with the field approach. Here the first PLS dimension was barely significant (cross-validated 0.88) and the second was not. In the score plot (Fig. 3) the resolution of the classes is seen to be moderate. We note that two of the sequences fall in the wrong domains in the plot, and hence are wrongly classified.

(4) DNA promoters ($L = 68$)

Promoters are specific DNA sequences that govern the binding of the sigma unit of the RNA polymerase holoenzyme (RNAP), thereby punctuating the onset of transcription. Jonsson et al. [2,26] recently investigated the relationship between the nucleic acid sequence of a set of $L = 68$ long promoters in *E. coli* and their in vivo promoter efficiency. Each position was coded by three descriptors, giving a position-based description of 195 variables (three positions were constant in the set). A PLS model with three significant components modelled $R^2 = 0.85$ of the variation of the *Y*-variable (promoter efficiency), with a cross-validated $Q^2 = 0.40$. This model was then used to predict how to construct sequences with even higher activity than the most active in the training set. The predicted "optimum" sequences were synthesized and indeed confirmed to be stronger promoters than the previously existing ones [2,27].

(5) Simulations of penta- and decapeptides ($L = 5$ and 10)

In the present examples the peptide and DNA sequence models based on either a position-based description or an auto- and cross-covariance (ACC) description give similar results, except for signal peptides where the ACC model is superior. This might be interpreted as that these two ways of translating three-way to two-way data are mathematically equivalent. To investigate this possibility, a limited set of simulations generating position-based data for artificial sets of (a) $N = 20$ penta- and (b) $N = 32$ decapeptides ($L = 5$ and 10) were made. The peptide sets were generated according to Plackett Burman or fractional factorial designs [18,19,21], respectively. An activity variable, *Y*, was generated as a linear model in the three *z*-scales times the L positions ($3 \times L$ variables) with uniformly $(-1, 1)$ randomly distributed coefficients. Normally distributed residuals (noise) with a standard deviation (S.D.) equal to 10% of the S.D. of Y was added to the *Y*-variable.

The generated data were then analysed analogously to examples 1, 2 and 4 with a position-based model and an ACC-based model. The position-

based model, as expected, adequately recovered the generated model (see Table 1). The cross-validation is seen to underestimate grossly the predictive power of the model for the data generated according to the fractional factorial designs (decapeptides), however. This phenomenon is well known and understood; the roundness and orthogonality of this design in X-space make regression and PLS models change direction substantially in X-space for each deletion of an observation in the cross-validation scheme, with the consequence that the noise in the data is overestimated.

For these simulations, the ACC model failed to give significant models according to the cross-validation. This result shows that the two ways of translating three-way into two-way data are not equivalent, and that the results of examples 1–4 indicate that periodicities and neighbour effects are important in these sets of biopolymer sequences.

TRANSLATING SEQUENCES AND PROCESSES TO DATA MATRICES

Once a "training set" of sequences is available, and their biological or/and physico-chemical properties have been measured, one needs to translate the actual sequences of the training set into numbers to be able to develop a relationship between sequence and activity or other properties.

Description of each position giving a three-way matrix

The translation of a sequence to quantitative variables is a difficult problem because it touches the very essence of chemistry and molecular biology, namely what structural features are important in a sequence and how to quantify them. There is the theoretical approach where a large number of quantum mechanical "indices" such as charges, partial molecular volumes and energy levels are calculated and used as descriptor variables.

Another approach is based on the so-called analogy principle, where one tries to use measurements on model systems to derive scales that are then used to describe to actual sequences. Along this line, three scales have been developed for the individual amino acid positions, the so-called z-scales [23,24]. These quantitative scales are developed from multivariate measurements in model systems and roughly correspond to hydrophilicity (z_1), steric bulk (z_2), and polarity/charge (z_3). Using these scales, one can hence translate a peptide sequence of length L into $3 \times L$ numbers. These can be arranged in a three-way matrix as in Fig. 1.

With a process time sequence, one usually measures values of the process variables such as pH, temperature or pressure, at various locations in the process set-up (feeders, reactors, distillation towers, etc.). These multivariate process measurements makes the data for each batch process be a matrix, and the data of a set of batch processes be a three-way matrix, here denoted **X**.

Output variables (responses, Y) measuring quality, yield, and cost are often measured less frequently, and in batch processes only at the end of the time sequence. Hence, the response matrix, **Y**, is often smaller than the process data matrix, **X**, and is often just a two-way matrix (batch × response).

When the sequences (polymer or time) have the same length and can be assumed to be alignable, these three-way matrices have data elements everywhere. For a set of sequences of different lengths, however, the three-way matrix will have parts without data, which is discussed further below.

Transforming the three-way matrix into a two-way data matrix

Depending on the degree of diversity between sequences or processes and assumptions about the "auto-correlation structure" in the sequences, the translation into the two-way matrix should be done in different ways.

In the modelling of time sequences of process data, it has long been recognized that the values of process variables at a time point t are not independent of the corresponding values at the time points $t-1$ and $t+1$. This has lead to the

use of time series models, where the correlations over time are taken into account [7]. One may argue that similar types of dependences might occur in biological sequences, and that time series tools here may have an interesting new area of application.

The simplest, and not always sufficient, way to account for dependences between consecutive observations is to use Fourier transforms, auto-correlation or auto-covariance transforms, or other suitable transforms of the data. Indeed, Van Heel [28] has recently shown that a qualitative 1–2 auto-correlation transform of protein sequences contains sufficient information to classify the proteins into known classes of great detail [28]. Here quantitative 1–2, 1–3 and 1–4 auto-correlation transforms of sequence descriptors will be used and compared, where possible, with the traditional position-based description.

Position / time aligned unfolding for similar sequences of almost the same length without neighbouring effects. When the biopolymer or process time sequences in a three-way matrix all are of the same length (or nearly so, see below), and a sequence position or process time point indeed corresponds to the same position or time point in the other sequences, one says that the sequences are alignable. One can then unfold the three-way table to give a two-way table as shown in Figs. 1b and 2b. This two-way matrix, **X**, can then be modelled and analysed according to one of the objectives stated above; overview, classification or quantitative relationship to Y.

When the sequences have almost the same length in chain or time, one can reach an approximate alignment by means of insertions of gaps at suitable positions in the shorter sequences. These gaps can then be parameterized by special values of the characterizing variables, or be left as missing data. The resulting three-way table, **X**, is then analysed in the same way.

This unfolding of **X**, followed by its analysis by a linear PC, classification, regression or PLS model, corresponds to the assumption that the positions of the sequences are truly independent, i.e., there are no neighbour–neighbour interactions. In process data this is a fairly unrealistic assumption, and one can use other ways of translating the three-way **X** into a two-way matrix as discussed below.

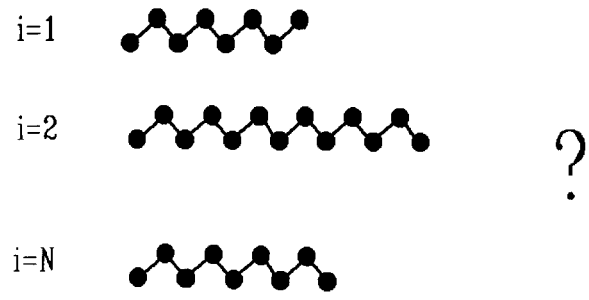


Fig. 4. Biopolymer sequences of different lengths cannot be directly quantified as in Fig. 1, because the sequences would give a two-way table with different numbers of columns, and the three-way table cannot be unfolded.

Transformations independent of alignment based on auto- and cross-covariance (ACC) structures. Whenever sequences in the same analysed set differ much in length (Figs. 4 and 5), one must use principles other than position-based to go from the three-way **X** to the two-way **X** table. However, also in the modelling of sequences of the same length one may wish to use a translation that takes neighbour effects, i.e., lack of independence between subsequent positions (time points), into account. This can be done by using ACC, and other transforms of each sequence, followed

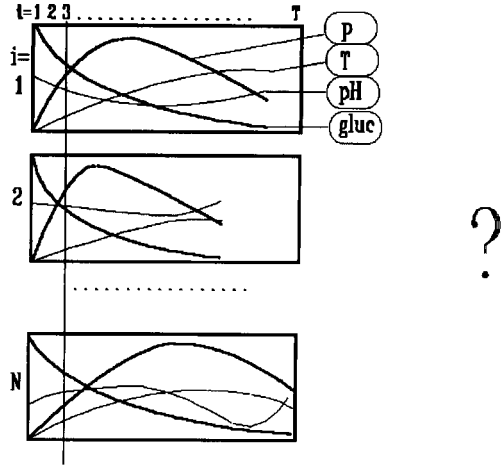


Fig. 5. Process time sequences of different lengths cannot be directly quantified as in Fig. 1, because the time sequences would give two-way table with different numbers of columns, and the three-way table cannot be unfolded.

TABLE 2

Example of the generation of auto- and cross-covariance functions (ACC) of a pentapeptide (RYLPT) described in each amino acid position by the three z-scales [23,24] [a]

Pos. (i)	AA (Amino acid)	z_1	z_2	z_3	$A_{11}(1)$ $z_1(i)$ $*z_1(i+1)$	$C_{12}(1)$ $z_1(i)$ $*z_2(i+1)$	$C_{13}(1)$ $z_1(i)$ $*z_3(i+1)$	$C_{21}(1)$ $z_2(i)$ $*z_1(i+1)$	$A_{22}(1)$ $z_2(i)$ $*z_2(i+1)$	$C_{23}(1)$ $z_2(i)$ $*z_3(i+1)$	$C_{31}(1)$ $z_3(i)$ $*z_1(i+1)$	$C_{32}(1)$ $z_3(i)$ $*z_2(i+1)$	$A_{33}(1)$ $z_3(i)$ $*z_3(i+1)$
1	Arg, R	1.13	−2.36	1.26	−4.0454	2.3391	−0.0452	8.4488	−4.8852	0.0944	−4.5108	2.6082	−0.0504
2	Tyr, Y	−3.58	2.07	−0.04	−13.4608	−4.1528	8.3056	7.7832	2.4012	−4.8024	−0.1504	−0.0464	0.0928
3	Lys, L	3.76	1.16	−2.32	0.0752	1.7672	11.5056	0.0232	0.5452	3.5496	−0.0464	−1.0904	−7.0992
4	Pro, P	0.02	0.47	3.06	0.011	−0.0446	−0.0298	0.2585	−1.0481	−0.7003	1.683	−6.8238	−4.5594
5	Thr, T	0.55	−2.23	−1.49									
Sum					−17.42	−0.0911	19.7362	16.5137	−2.9869	−1.8587	−3.0246	−5.3524	−11.6162
Sum/4		ACC(1)			−4.355	−0.023	4.934	4.128	−0.747	−0.465	−0.756	−1.338	−2.904
Pos. (i)	AA (Amino acid)	z_1	z_2	z_3	$A_{11}(2)$ $z_1(i)$ $*z_1(i+2)$	$C_{12}(2)$ $z_1(i)$ $*z_2(i+2)$	$C_{13}(2)$ $z_1(i)$ $*z_3(i+2)$	$C_{21}(2)$ $z_2(i)$ $*z_1(i+2)$	$A_{22}(2)$ $z_2(i)$ $*z_2(i+2)$	$C_{23}(2)$ $z_2(i)$ $*z_3(i+2)$	$C_{31}(2)$ $z_3(i)$ $*z_1(i+2)$	$C_{32}(2)$ $z_3(i)$ $*z_2(i+2)$	$A_{33}(2)$ $z_3(i)$ $*z_3(i+2)$
1	Arg, R	1.13	−2.36	1.26	4.2488	1.3108	−2.6216	−8.8736	−2.7376	5.4752	4.7376	1.4616	−2.9232
2	Tyr, Y	−3.58	2.07	−0.04	−0.0716	−1.6826	−10.9548	0.0414	0.9729	6.3342	−0.0008	−0.0188	−0.1224
3	Lys, L	3.76	1.16	−2.32	2.068	−8.3848	−5.6024	0.638	−2.5868	−1.7284	−1.276	5.1736	3.4568
4	Pro, P	0.02	0.47	3.06									
5	Thr, T	0.55	−2.23	−1.49									
Sum					6.2452	−8.7566	−19.1788	−8.1942	−4.3515	10.081	3.4608	6.6164	0.4112
Sum/3		ACC(2)			2.082	−2.919	−6.393	−2.731	−1.451	3.360	1.154	2.205	0.137

[a] The calculations of the auto and cross covariance functions of lag d, $A(d)$ and $C(d)$ respectively, are described in the text. For lag $d = 1$, the contributions to an ACC, for example $C_{12}(1)$, are formed by multiplying the z_1-value of position i by the z_2-value of position $i+1$ for $i = 1, 2, 3$, and 4. Thereafter these four terms are summed and divided by 4 to give the average = $C_{12}(1)$.

by the digitization and unfolding of suitable parts of these transforms.

In pioneering work, Van Heel [28] transformed a large number of protein sequences into nearest neighbour amino acid frequency histograms. In each sequence, he counted how many times the pair Ala–Ala appears, the pair Ala–Asn, the pair Ala–Asp, etc., until the pair Tyr–Tyr. This gave 400 relative frequencies, one for each of the 20×20 amino acid pairs. A characterization based on these 400 values can be used as variables to characterize any peptide sequence, independent of its length or alignment. Van Heel then showed that these variables contain sufficient information about similarities and dissimilarities of the proteins to classify them into known classes with very fine detail.

The advantages of Van Heel's sequence characterization are that it is general and independent of alignment, and that it accounts for nearest neighbour interactions. It has some drawbacks, however, in that it is qualitative (all pairs of amino acids are equivalent) and difficult to interpret.

Here a simplification and quantification of Van Heel's scheme is proposed by using first the three z-scales (and a fourth for D-/L- if warranted), and then calculating the ACC functions of this description along the sequence. With three z-scales this gives nine nearest neighbour (lag 1) ACCs, nine next nearest neighbour (lag 2), etc., because one computes also the ACCs between z_1 and z_2, between z_1 and z_3, etc. To simplify further the description and calculations, the auto- and cross-covariance functions are used instead of auto- and cross-correlations (see below). The abbreviation ACC will be used, somewhat loosely, for both auto- and cross-correlations and covariances, which is acceptable as they are almost identical after scaling.

Table 2 gives an example of the calculation of the lag 1 and lag 2 ACCs of a pentapeptide as an illustration.

As pointed out by Van Heel [28], it may be warranted to compute the ACCs separately for, say, the beginning of a sequence, the middle of a sequence and the end of a sequence, particularly when the sequences are long. However, in all examples including the signal peptides a single ACC function is used over the whole sequence. Van Heel's proposal will be investigated in the future in connection with a more systematic study of sequence modelling. To complement the ACC description, averages of the z-values for different parts of the sequences, say the beginning, middle and end, may provide additional variables of possible modelling value.

Auto- and cross-correlation and covariance (ACC)-based transformations. The lag d auto-correlation function of a descriptor x (say z_1) over a sequence of length L is (the subscript i is the sequence position index running from 1 to L) [7] is given by

$$A(d) = \sum_i (x_i - \bar{x})(x_{i+d} - \bar{x}) / \sum_i (x_i - \bar{x})^2$$

The corresponding auto-covariance function is [7]

$$A(d) = \sum_i (x_i - \bar{x})(x_{i+d} - \bar{x})/L$$

A lagged cross-covariance function between two descriptors x and u, e.g., $x = z_1$ and $u = z_2$, is analogously defined as [7]

$$C_{xu}(d) = \sum_i (x_i - \bar{x})(u_{i+d} - \bar{u})/L$$

The summations are done from 1 to $L - d$ because of terms such as x_{i+d}. Also, it is important to note that ACC_{xu} functions differ from ACC_{ux}. Table 2 gives an example of the calculation of the lag 1 and lag 2 ACC functions of a pentapeptide.

In the calculation of ACC functions of the biopolymer sequences centred terms were not used as above, i.e., subtracting $\bar{x}$, etc., but rather used uncentered terms corresponding to using $\bar{x} = 0$, because the z-scales are already centred over all amino acids in their derivation.

MULTIVARIATE MODELLING AND ANALYSIS

Once a set of sequences have been translated into a two-way data matrix, standard methods of multivariate analysis such as PCA, and PLS are used to develop the models relating sequence to properties (or class) [1,2,8,9,14–17]. With both polymer and process time sequences, evolving

factor analysis (EFA) [29] and evolving PLS [30] may be interesting alternatives aimed at an early diagnosis of sequence "type" or class, and possibly capable of recognizing "clean" parts of a sequence corresponding to a single action mechanism.

In this study PLS modelling and PLS discriminant analysis [8,14–17] were used in all examples. The two-way data matrices were all auto-scaled to unit column variance before the analysis, and cross-validation was used to ensure predictive significance of the models. SIMCA-4R software [31] was used for the computations.

RESULTS

Table 1 summarizes the results of the simulations and the four examples. Only example 3 will be discussed in more detail because it is the most general with sequences of different lengths, and because the ACC model gives such good results.

Simulations

The main result is that the position-based models and ACC-based models are not equivalent. Data generated according to a position-based model are, in general, very poorly modelled by an ACC model unless periodic structures are explicitly generated.

Examples 1, 2 and 4

The interesting result of these examples is that the ACC models give the same fit (R^2) and slightly better cross-validated predictive power than the position-based models, even for the very short peptides (see Table 1). This indicates that the positions in these sequences are not independent, which may have consequences for the design and modifications of biopolymer sequences as discussed further below.

Example 3, signal peptides

Each signal sequence was first multi-positionally described by the three z-scales. Then the ACC model was used to generate 54 variables. Thus all possible cross-terms were generated between amino acid position i and position $i+d$, and where d is between 1 and 6. ($3 \times 6 \times 3 =$ scales $\times$ lagged positions $\times$ scales $= 54$). These variables are labeled a to f for lags 1 to 6, followed by the two subscripts of the cross-covari-

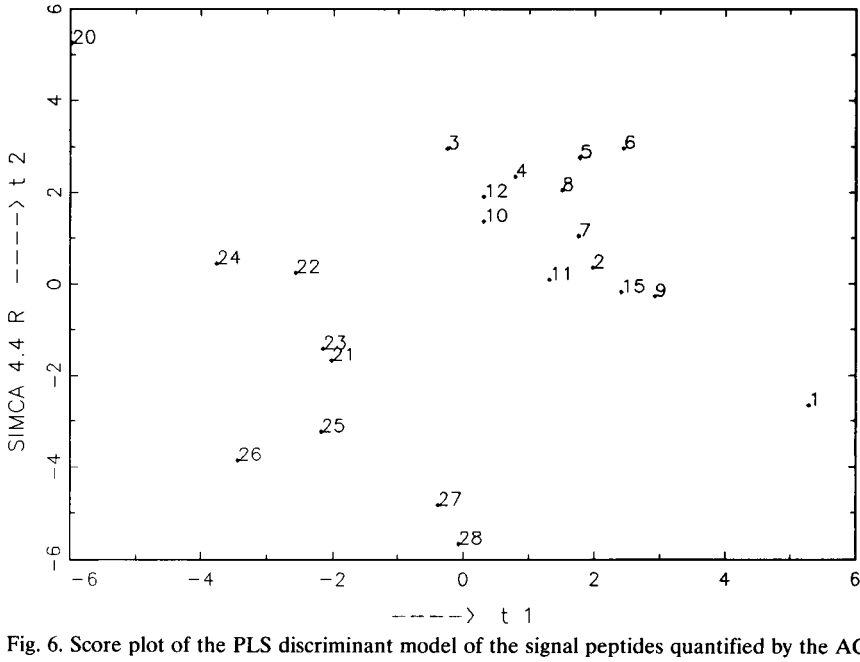


Fig. 6. Score plot of the PLS discriminant model of the signal peptides quantified by the ACC functions of the z-scaled. Nos. 1–15 are periplasmic space signal peptides and Nos. 20–28 are outer membrane signal peptides.

ated z-scales. Hence d_{31}, for example, denotes the lag 4 cross-covariance between z_3 and z_1.

A dummy variable with the value $y = 1$ was used for the periplasmic space proteins and $y = -1$ for the outer membrane proteins. The PLS discriminant analysis, based on auto-scaled data, showed that two components were significant according to cross-validation, with $Q^2 = 0.63$. The two PLS dimensions model 64% and 26%, respectively of the variation in y.

A score plot of the two significant components (t_2/t_1) is given in Fig. 6 and the corresponding loading plot in Fig. 7. The score plot reveals a very good separation between the classes, with all training sequences correctly classified. The PLS weight plot (Fig. 7) shows that many of the ACC functions contribute with information to the class separation, dominated by the terms of gap 4, z_3, z_1, and z_1, z_1 (d_{31} and d_{11}) and the gaps 3, and 6, z_2, z_1 terms (c_{21} and f_{21}).

The negative signs of the d_{31} terms show that in signal peptides of outer membrane proteins d_{31} is high (because $y = -1$ for these). This means that in these sequences there is a strong tendency to have lipophilic amino acids such as Val, Leu and Phe (negative z_1 values), situated four positions after amino acids with positive z_3 values (Pro, Cys, Asp), or hydrophilic amino acids with positive z_1 values (e.g., Arg, Asn, Asp, Glu) four positions after amino acids with negative z_3 values (e.g., Lys, Arg).

Signal peptides of periplasmic proteins, in contrast, are characterized by positive d_{11}, c_{21}, and f_{21} values, i.e., amino acids separated by a gap of 4 have a tendency of being either both hydrophilic or both lipophilic (d_{11}), and there are positive correlations between large size and hydrophilicity, or small size and lipophilicity, respectively, 3 and 6 positions later (c_{21} and f_{21}).

These coefficients can possibly be interpreted in terms of helix-making or helix-breaking patterns, and these possibilities will be pursued later in a more detailed investigation of a larger set of signal peptides. It is worth noting that the strongest patterns in both classes involve lagged cross-covariances between different properties which easily escape traditional sequence analyses of one separate property at a time.

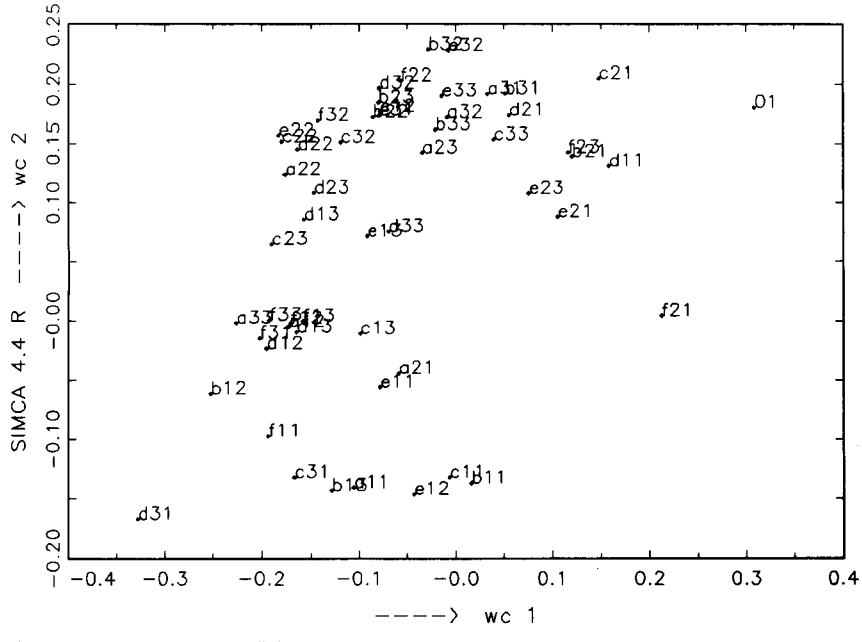


Fig. 7. Plot of the PLS coefficients w and c of the two dimensions of the ACC PLS discriminant model of the signal peptides. The point 01 indicates the position of the dummy y-variable. The model is seen to be dominated by the terms d_{31} (lag 4, z_3, z_1 cross correlation), c_{21}, d_{11}, and f_{21} terms.

ADVANTAGES AND DISADVANTAGES WITH ACC-BASED MODELS

ACC modelling has some clear advantages in that the description and modelling of a set of polymer or time sequences become independent of alignment. Whenever a set of sequences have widely different lengths, as in the signal peptide example, position-based models requiring alignment are difficult or impossible. Moreover, consistent dependences between neighbouring sequence positions can be modelled.

However, a clear difficulty with ACC models, which they share with any multivariate models where interactions between variables are important, is the interpretation and understanding of the model. This is very much an educational problem; we are taught to see the world in terms of independent variables and individual causal factors, even when it is obvious that many natural systems display strong interactions, multiple causes and lack of independence. Auto- and cross-covariance patterns correspond to repeated interactions of the same kind over the sequence, and are hence difficult to comprehend.

This difficulty of interpretation corresponds to difficulties in predicting sequences with higher or lower values of the response variables and constructing a typical sequence belonging to a certain class, say a periplasmic signal peptide. Only systematic simulations, starting with a given sequence and making multiple simulated changes, and inserting each modified sequence into the model, will result in sequences predicted to be "optimal" for a given purpose. These multiple simulated changes of the starting sequence must, of course, be made according to an appropriate statistical design [18,19,21], in order for the results to be reliable.

DISCUSSION AND CONCLUSION

There are many ways to quantify biopolymer and process data time sequences. The use of theoretical calculations by molecular mechanics and quantum chemistry (molecular modelling) is always possible, but has the drawback of being complicated and time consuming, especially for longer sequences. The molecular modelling of sets of sequences of different length introduces problems of equivalence; it is not certain that theoretically derived parameters are directly comparable for sequences of different length.

Process modelling has similar problems: knowing what to measure, where and how often requires great insight and experience.

The examples in this paper demonstrate that the simple and straightforward quantification of biopolymer sequences by z-scales that are derived from physico-chemical properties of free amino acids (or nucleotides) contains much information about the properties of the sequences.

Position-based modelling of the resulting data works as long as sequences are of the same length and similar so that indeed a position in different sequences has the same chemical and biological significance, and as long as sequence periodicities lack biological significance; and analogously for process time sequences. Interactions between different parts of the sequence could be modelled by interaction terms between variables at different positions. With a sequence of some length, the number of possible interactions becomes very large, however; in a decapeptide there are $10 \times 9/2$ possible position–position interactions, which, when multiplied by 9 for all possible z–z cross-interactions, becomes 405.

For sets of biopolymer sequences of different length, the further preprocessing of the sequence characterizing data by means of ACC (auto- and cross-correlation or -covariance) transformations gives alignment-independent and general descriptions of the sequences, which also seem to preserve substantial parts of the information in the data. Averages of the z-values for different parts of the sequences may provide complementary information regarding the "average" structure of sequences, particularly in classification problems.

Similarly, multivariate ACC modelling (including cross-correlations) of sets of process data time sequences is an interesting and simple alternative to process time series modelling, e.g., of batch processes. Batch processes often differ in length because the time of "completion" is influenced by changes in process variables such as pH, tem-

perature and pressure during the course of the process. Fermentation processes in biotechnology provide typical examples.

The fact that ACC modelling of biopolymers works so well indicates that nature indeed assigns significance to "periodic" patterns. The biochemical interpretation of this finding is far from clear and much more experimental validation and experience of the ACC models is necessary. Some periodicities related to helices of peptide sequences are known and possible to recognize, but the extent of other periodic patterns in biopolymer sequences is little understood. The possibility that the recognition of sequences by biological "receptors" may be based on auto-correlation patterns is most interesting, however, and may have profound consequences for biochemical theory. The results of van Heel [28] and, to some extent the present investigation raise this possibility.

This line of investigation is continuing, looking at a wider range of data and at such questions as the use of a single or several ACC structures, possible relationships between ACC structures and protein folding and other secondary and tertiary structures, i.e., helices, sheets, bends, and motifs involving their combinations. The use of ACC structures with QSARs of ordinary organic molecules that are not sequences is also an intriguing possibility. In addition, the use of these models in analytical chemistry to predict chromatographic separation and other properties of peptides and nucleotides is of great practical interest.

Support from the Swedish Natural Science Research Council (NFR), the Swedish Board for Technical Development (STU–NUTEC) and the Centre for Environmental Research (CMF) is gratefully acknowledged.

REFERENCES

1 W.J. Dunn and S. Wold, in C. Hansch and C. Ramsden (Eds.), Comprehensive Medicinal Chemistry, Vol. 4, Pergamon Press, Oxford, 1990, Chap. 22.3.
2 J. Jonsson, Thesis, Research Group for Chemometrics, Umeå University, Umeå, 1992.
3 A. Lorber and B.R. Kowalski, J. Chemometr., 2 (1988) 67.
4 B. Skagerberg, J.F. MacGregor and C. Kiparissides, Chemometr. Intell. Lab. Syst., 14 (1992) 341.
5 J.V. Kresta, J.F. MacGregor and T.E. Marlin, Can. J. Chem. Eng., 69 (1991) 35.
6 P. Nomikos and J.F. MacGregor, in NATO ASI Symposium on Batch Processes, Antalia, Turkey, May 1992, Springer, New York, 1993.
7 G.E.P. Box and G.M. Jenkins, Time Series Analysis, Holden-Day, Oakland, CA, 2nd edn., 1976.
8 J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
9 S. Wold, K. Esbensen and P. Geladi, Chemometr. Intell. Lab. Syst., 2 (1987) 37.
10 Integrated Process Intelligence (IPI), Manual of SIMCA-P, Version 1.0, Umetri, Umeå, and ABB–AUT, Luleå, 1992.
11 C. Albano, W.J. Dunn, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.
12 K.V. Mardia, J.T. Kent and J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.
13 S. Wold, Pattern Recognition, 8 (1976) 127.
14 M. Sjöström, S. Wold and B. Söderström, in E.S. Gelsema and L.N. Kanal (Eds.), Pattern Recognition in Practice, II, Elsevier, Amsterdam, 1986, p. 486.
15 L. Ståhle and S. Wold, J. Chemometr., 1 (1987) 185.
16 H. Wold, in K.-G. Jöreskog and H. Wold (Eds.), Systems Under Indirect Observation, Vol. II, North-Holland, Amsterdam, 1982, Chap. 1.
17 A. Höskuldsson, J. Chemometr., 2 (1988) 211.
18 G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, Wiley, New York, 1978.
19 E. Morgan, Chemometrics: Experimental Design, ACOL, London, and Wiley, New York, 1991.
20 N.R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 2nd edn., 1978.
21 S. Hellberg, M. Sjöström, B. Skagerberg, C. Wikström and S. Wold, Acta Pharm. Jugosl., 37 (1987) 53.
22 L. Eriksson, J. Jonsson, S. Hellberg, F. Lindgren, B. Skagerberg, M. Sjöström and S. Wold, Acta Chem. Scand., 44 (1990) 50.
23 S. Hellberg, Thesis, Research Group for Chemometrics, Umeå University, Umeå, 1986.
24 J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström and S. Wold, Quant. Struct. Activ. Relat., 8 (1989) 204.
25 M. Sjöström, S. Wold, Å. Wieslander and L. Rilfors, EMBO J., 6 (1987) 823.
26 J. Jonsson, L. Eriksson, S. Hellberg, F. Lindgren, M. Sjöström and S. Wold, Acta Chem. Scand., 45 (1991) 186.
27 J. Jonsson, T. Norberg, L. Carlsson, C. Gustafsson and S. Wold, Nucl. Acid Res., (1993) in press.
28 M. van Heel, J. Mol. Biol., 220 (1991) 877.
29 H. Gampp, M. Maeder, C.J. Meyer and A.Z. Zuberbühler, Talanta, 32 (1985) 1133.
30 K. Helland, H.E. Berntsen, O.S. Borgen and H. Martens, Chemometr. Intell. Lab. Syst., 14 (1992) 129.
31 Simca-R 4.4, Manual, Umetri, Umeå, 1992.

Analytica Chimica Acta, 277 (1993) 255–266
Elsevier Science Publishers B.V., Amsterdam

Superfund site characterization using non-parametric variogram modeling

Ashok K. Singh
Department of Mathematics, University of Nevada, Las Vegas, NV 89154 (USA)

Malwane M.A. Ananda
Department of Mathematics, University of Nevada, Las Vegas, NV 89154 (USA)

Allen R. Sparks
U.S. Environmental Protection Agency, Las Vegas, NV 89154 (USA)

(Received 3rd September 1992)

Abstract

The spatial interpolation method of Kriging, originally developed for estimation of ore reserves in mining and prediction in meteorology, has become an important geostatical tool for assessment of contamination in Superfund site characterization. Kriging uses an estimated spatial covariance matrix, or alternatively the variogram, to find the best linear unbiased estimate (BLUE) of contamination at an unsampled location. The modeling of the variogram (also called semi-variogram) is therefore a necessary step that must be performed before using Kriging for spatial interpolation. In this paper, we propose and investigate a simple non-parametric procedure based on ranks for variogram modeling. Some simulated examples are used to illustrate the proposed procedure. The proposed procedure is compared with the usual method of Ordinary Kriging for the Billings Smelter Superfund site data.

Keywords: Superfund site characterization

The characterization of a Superfund site for one or more contaminants is an important issue in making remediation decisions. Kriging, a geostatistical method of spatial estimation, is the commonly used tool for estimation of the amount of contaminant present at an unsampled location [1]. Kriging was originally developed for estimation of ore reserves in mining by Matheron [2] in France. Very similar methods were developed for meteorological predictions by Gandin [3] in the Soviet Union.

Correspondence to: A.K. Singh, Department of Mathematics, University of Nevada, Las Vegas, NV 89154 (USA).

The method of Kriging produces an optimal estimator in the sense of best linear unbiased estimation if the model for spatial dependence for the contaminant of concern is known, and a near-optimal estimator when the model for spatial dependence is to be estimated from the sampled data. In any Superfund site characterization investigation, the model for spatial dependence is completely unknown, and must be first estimated from the collected data set. Most of the classical statistical estimation methods start with the assumption of independence of observations, which is clearly violated in a typical spatial data set. Because of the lack of independence, a spatial data set of n observations, instead of being a

'sample of size n', is in fact a sample of size 1 from a multivariate distribution. This necessitates imposing some assumptions on the model for spatial dependence so that it can be estimated from a sample of size 1. Different set of assumptions lead to different forms of Kriging. In this paper, we will only be concerned with Ordinary Kriging (OK), in which first and second order stationarity of data are assumed to obtain estimates of the contaminant of concern at an unsampled location, and also the Kriging standard deviation (Ksd) of the estimates. Symmetry of the distribution of data is an implicit assumption for obtaining the Kriging estimates and their Ksds. If the goal is to form confidence intervals, then normality of the distribution is also required. A commonly used practice in geostatistics is to use the logarithm (or another suitable transformation) of the data to obtain approximate normality of the distribution. One problem with this approach is that an estimator which is a linear function of the transformed variable is not a linear function of the original variable. In the usual approach, one carries out all the steps of Kriging on the transformed variable, and then uses appropriate formulas to back-transform the Kriging estimates and also compute the Ksds of the back-transformed estimates.

In the present paper, a slightly different approach is taken. Here we use the rank transformation of data only for variogram modeling. The rank transformation has been used in the standard literature to derive non-parametric versions of parametric procedures [4]. The variogram based on ranks can be used to Krige the rank-transformed data, then the Kriged ranks can be back-transformed to obtain Kriged estimates of the original variable. One problem with this approach is that the Ksds of ranks are not easy to back-transform. This, however, is the subject of a future paper of ours. In this paper, we show how the variogram on ranks can be scaled to obtain a robust estimator of the variogram of the original variable. Other robust estimators of the variogram exist in the geostatical literature [5,6]. One advantage of our approach is its simplicity. No new software needs to be developed to use our method. The software package Geo-EAS, developed by the USEPA, can be used to perform the variogram modeling on ranks, and the program TRANS can then be used to compute the Ksds of the estimates.

Below, the estimator of the variogram model for the original variable based on the estimated variogram model for ranks is given and several examples generated by conditional simulation are used to illustrate our procedure. The data from the Billings Smelter Superefund site in New Mexico is Kriged using the standard OK method and also our method. Both methods yield very similar results.

ESTIMATOR OF THE VARIOGRAM BASED ON RANKS

Following the notations and terminology of Cressie [7], let $\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbf{R}^d\}$ be a real-valued stochastic process defined on a domain D of the d-dimensional space $\mathbf{R}^d$. It is assumed in parametric Kriging that the process $Z(s)$ is *intrinsically stationary*, i.e., it satisfies the following two conditions:

$$E[Z(\mathbf{s})] = \mu, \ \mathbf{s} \in D \quad (1)$$

$$\text{var}[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})] = 2\gamma(\mathbf{h}), \ \mathbf{s}, \mathbf{s}+\mathbf{h} \in D \quad (2)$$

In this paper, we will restrict attention to the case of *isotropic semivariogram*, in which case $\gamma(\mathbf{h}) = \gamma(\|\mathbf{h}\|)$, where $\|\mathbf{h}\|$ denotes the length of the vector $\mathbf{h}$.

Let $\{Z(\mathbf{s}_i): i = 1, 2, \ldots, n\}$ be n realizations of the process at n different locations in the set D. The problem is to predict $Z(\mathbf{s}_0)$ at some unsampled locations $\mathbf{s}_0$. The method of kriging finds the linear unbiased estimator with minimum variance. Using Lagrangian multiplier m for the unbiasedness condition, one obtains an $(n+1)\times(n+1)$ system of linear equations

$$\boldsymbol{\Gamma}\boldsymbol{\lambda} = \boldsymbol{\gamma} \quad (3)$$

where $\boldsymbol{\Gamma}$ is a symmetric $(n+1)\times(n+1)$ real matrix defined by

$$\begin{aligned} \Gamma_{ij} &= \gamma(\mathbf{s}_i - \mathbf{s}_j), & i &= 1, \ldots, n, \ j = 1, \ldots, n \\ &= 1, & i &= n+1, \ j = 1, \ldots, n \\ &= 0, & i &= n+1, \ j = n+1 \end{aligned}$$

and

$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n, m)'$,

$\boldsymbol{\gamma} = \gamma(\mathbf{s}_1, \ldots, \mathbf{s}_n, 1)'$

The BLUE is given by

$$\hat{Z}(\mathbf{s}_0) = \sum_1^n \lambda_i Z(\mathbf{s}_i) \tag{4}$$

and its variance, called kriging variance, is given by

$$E\left[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)\right]^2 = \sum_1^n \lambda_i \gamma(\mathbf{s}_i - \mathbf{s}_0) + m \tag{5}$$

where $(\lambda_1, \ldots, \lambda_n, m)'$ is the solution of the linear system of Eqn. 3. The isotropic semivariogram function $\gamma(.)$ is modelled from the given data.

Under the assumptions of first order stationarity and the intrinsic hypothesis, the variogram and covariance functions have the following relationship:

$$\gamma(Z_j, Z_j) = \sigma^2 - \mathrm{COV}(Z_j, Z_j) \tag{6}$$

which can be expressed as

$$\gamma(Z_j, Z_j) = \sigma^2\left[1 - \mathrm{CORR}(Z_j, Z_j)\right] \tag{7}$$

A non-parametric estimator of the correlation function $\mathrm{CORR}(Z_j, Z_j)$ is the rank correlation function $\mathrm{CORR}(R_j, R_j)$, which can be used to form an estimator of the variogram $\gamma(Z_j, Z_j)$ for the original variable as follows:

$$\hat{\gamma} = \mathrm{VAR}(R_j)\left[1 - \mathrm{CORR}(R_j, R_j)\right]\frac{\sigma^2}{\mathrm{VAR}(R_j)} \tag{8}$$

which, by Eqn. 6, is simply

$$\hat{\gamma}(Z_j, Z_j) = \mathrm{CF}\gamma(R_j, R_j) \tag{9}$$

where CF (correction factor) is a constant given by:

$$\mathrm{CF} = \frac{\sigma^2}{\mathrm{VAR}(R_j)} \tag{10}$$

Our method of variogram modeling and Kriging can now be described in a few steps:

(1) Sort the data set $\{Z_i, i = 1, 2, \ldots, n\}$ and compute the mid-rank of Z_i for each $i = 1, 2, \ldots, n$.

(2) Replace the data Z_i by its mid-rank R_i.

(3) Compute a variogram model for the ranks.

(4) Multiply the variogram model obtained in step 3 by the constant CF given by Eqn. 10 to obtain a non-parametric estimator of the variogram for Z values.

(5) Compute Kriging estimates from the variogram model of step 4.

All of these step can be carried out by using the software package GEO-EAS.

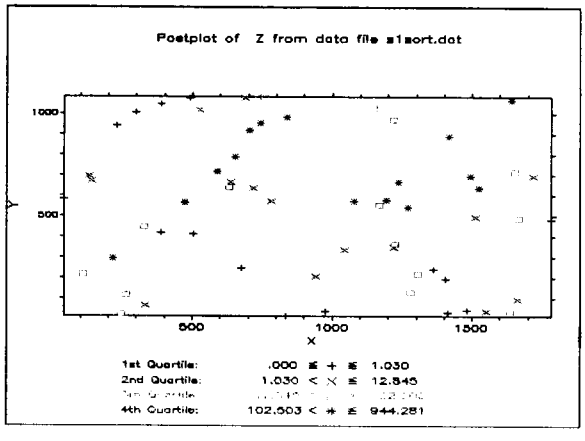
Since the non-parametric estimator for the variogram for Z differs from the variogram for ranks only by a scale factor, we can alternatively perform Kriging using the variogram for ranks, and then multiply the resulting Kriging variance by the constant CF of Eqn. 5 to obtain the correct Kriging variance of the estimator. The Kriging estimates, of course, are not affected by the missing scale factor CF. This is the route taken below.

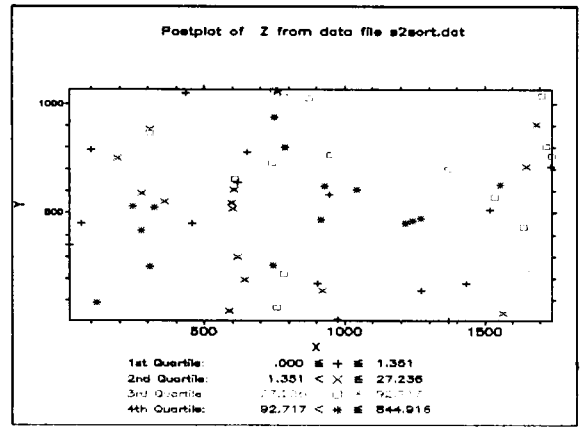
CONDITIONAL SIMULATION EXAMPLES

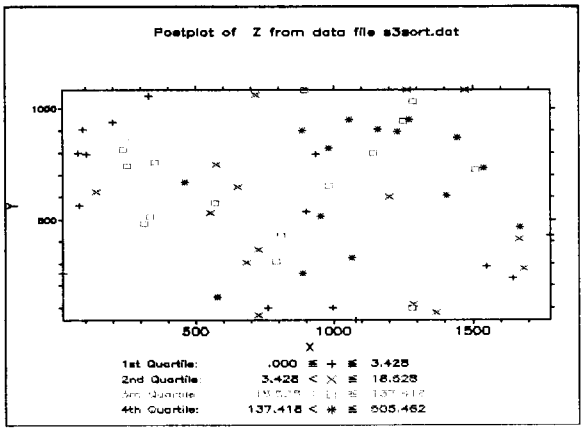
A test case, for which the true values of the field are known, is the well-known Walker Lake data set [8]. The Walker Lake data set was derived from digital elevation data from the Walker Lake area near the California–Nevada border. A digital elevation model of the National Cartographic Information Center (NCIC), containing elevations for 1.95 million points on a regular grid, was transformed into a data set of 78000 points on a 260×360 rectangular grid.

In a study of the amount of subjectivity present in the use of Kriging, Englund [9] used two rectangular subsets of the Walker Lake data set — Area A (NE portion of the Walker Lake), and Area B (SE portion of the Walker Lake), for his experiment. We have drawn several samples of size 60 each from the Area A for our examples. Results of four of these examples are presented here. In each of these four examples, we have only fitted the omnidirectional variogram to keep things simple; no attempt has been made to find the best variogram model in each case. The software package GEO-EAS was used for all of the computations in this paper.

Figure 1 shows a shaded map of the Area A







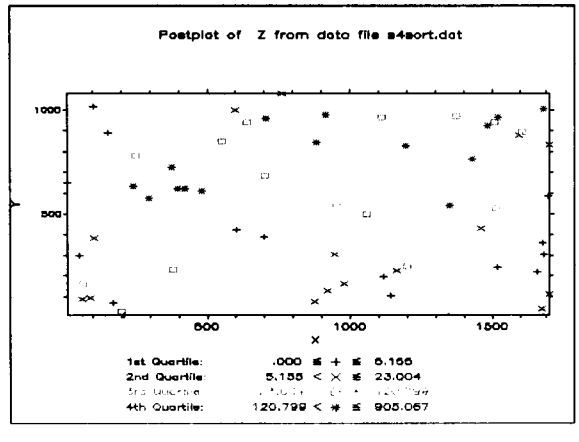


Fig. 1. Shaded map of the Area A from Walker Lake data set, and map of sampling locations of the four data sets of size $n = 60$.

data set, and maps of the 60 sampling locations for each of the four sample data sets from Area A. Figures 2, 4, 6, and 8 show the results of variogram modeling on the original variables Z (first plot), cross-validation (second and third plots), and Kriging on Z for the four data sets, respectively. Figures 3, 5, 7, and 9 show the corresponding results for the method proposed in this paper. It appears from these figures that, except for the data set 3, the variogram based on ranks is at least as good as the variogram based on Z itself. The variogram based on ranks has the additional advantage of being robust.

As mentioned, the Ksds of the estimates produced by Kriging with the rank-based variogram model are off by a scale factor. We must multiply the resulting Ksd by the square-root of the correction factor CF given by Eqn. 10. For example, for the data set 1 from Area A, CF equals $51\,948/299.62 = 173.4$, and hence we must multiply the Ksds resulting from Kriging with the rank-based variogram by its square-root $173.5^{0.5} = 13.2$ to obtain the correct Ksds. This is the reason for the large discrepancy between the values shown in the histogram plots (third plot from top) of Figs. 2 and 3. Since these histograms are for the standardized errors (estimate − Z)/Ksd obtained from cross-validation, the values shown in the third plot of Fig. 3 must be divided by 13.2. For the first example data set from Area A, Table 1 shows the correct summary statistics for the standardized errors from cross-validation for the proposed method, along with the corresponding values obtained from the usual OK.

It appears from Figs. 2–9 that the semi-variograms based on Z are different from those based on ranks, but the Kriged maps (last plot of these figures) are quite similar, for each of the four example data sets from Area A.

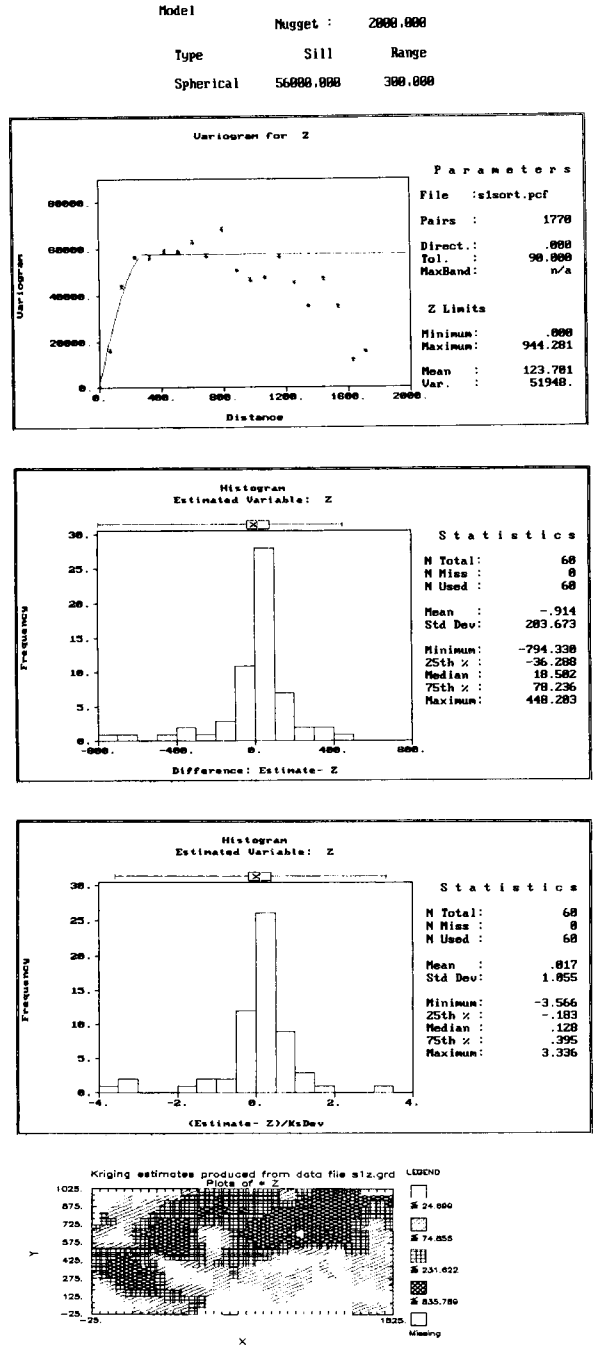


Fig. 2. Results of variogram modeling on Z, cross-validation, and Kriging on Z for the first data set from Area A.

SITE CHARACTERIZATION OF THE CAL WEST METALS NPL SITE

Cal West Metals is located one mile north of Lemitar and about nine miles north of Socorro, New Mexico. This facility was operated as a battery recycler and a secondary lead smelter between 1979 and 1981. In addition, research was

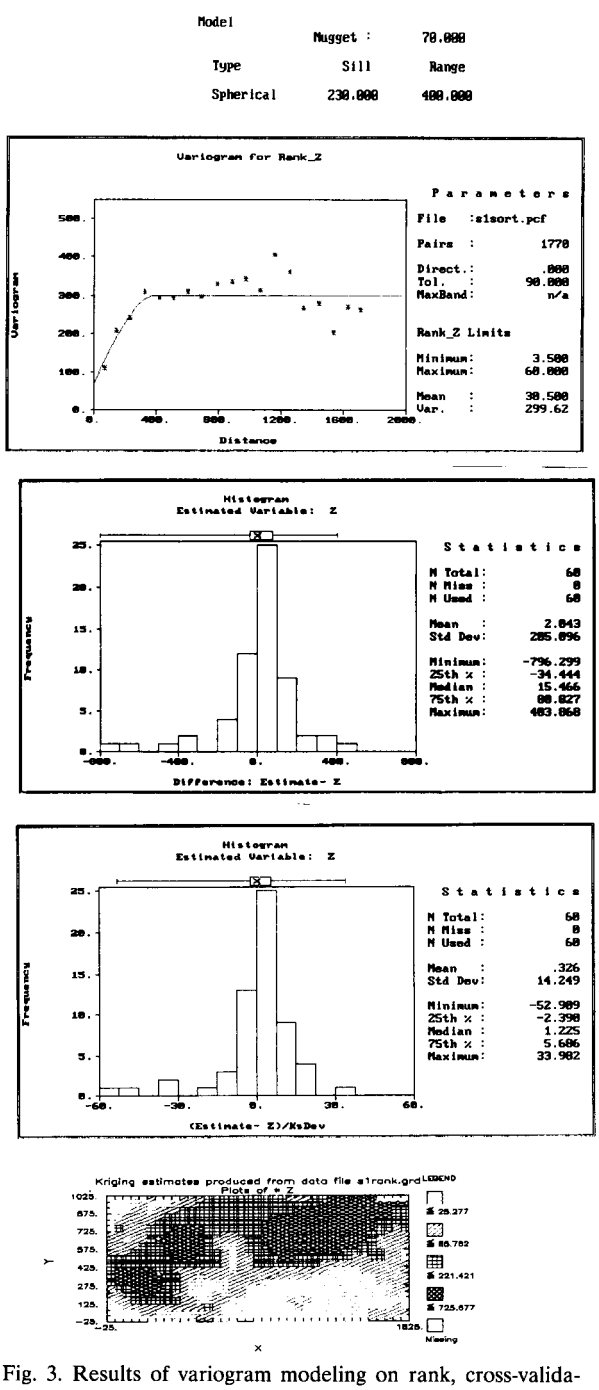


Fig. 3. Results of variogram modeling on rank, cross-validation, and Kriging on Z for the first data set from Area A.

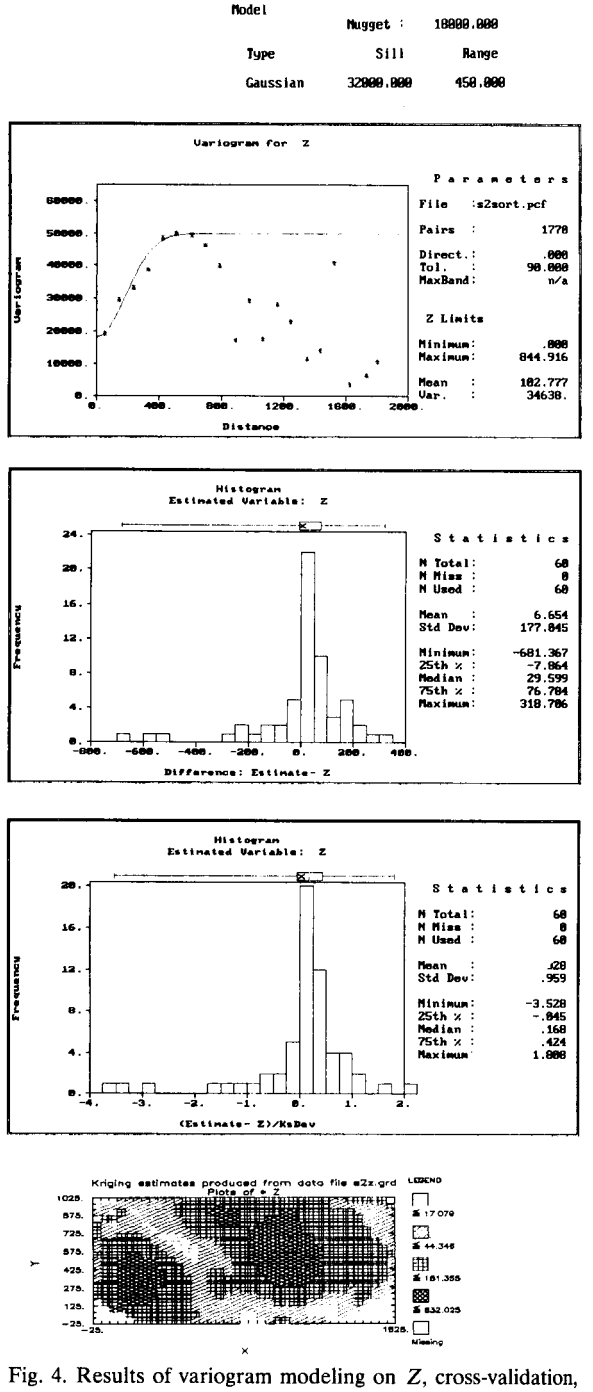


Fig. 4. Results of variogram modeling on Z, cross-validation, and Kriging on Z for the second data set from Area A.

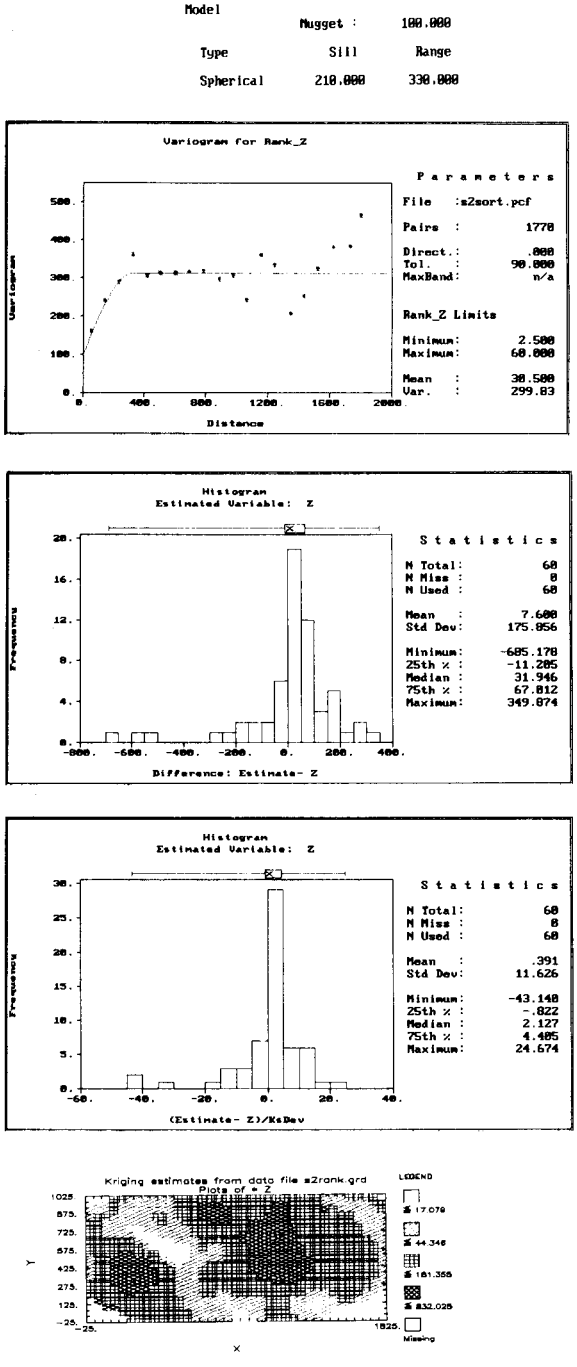


Fig. 5. Results of variogram modeling on rank, cross-validation, and Kriging on Z for the second data set from Area A.

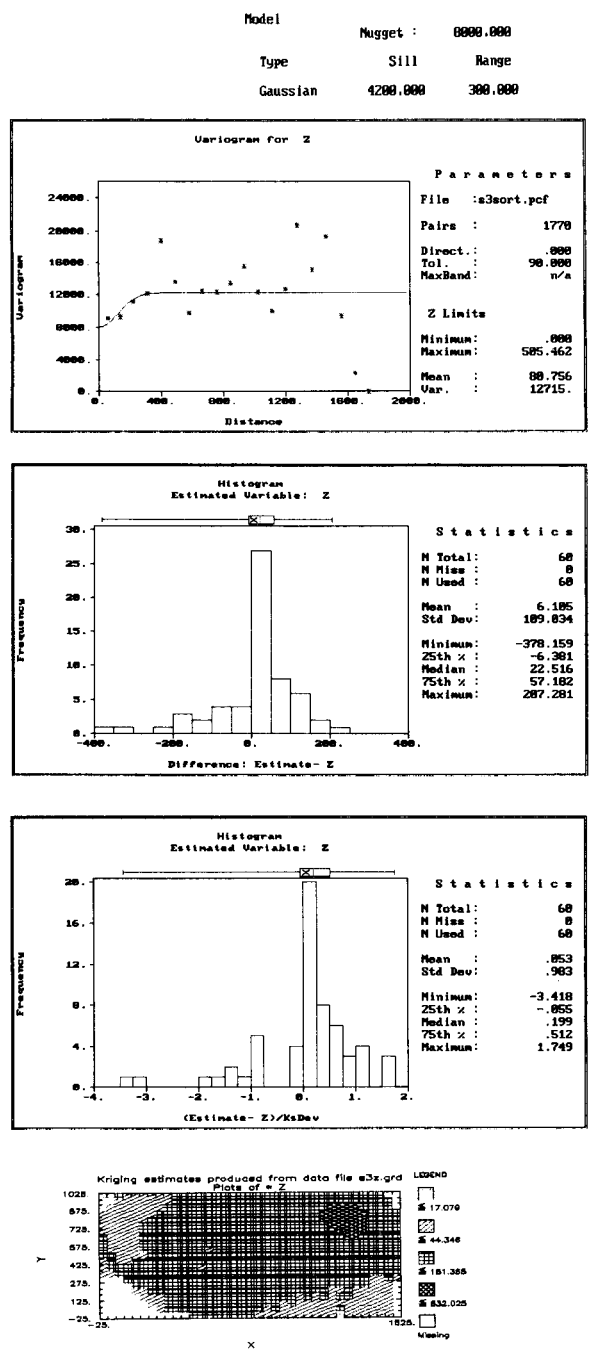


Fig. 6. Results of variogram modeling on Z, cross-validation, and Kriging on Z for the third data set from Area A.

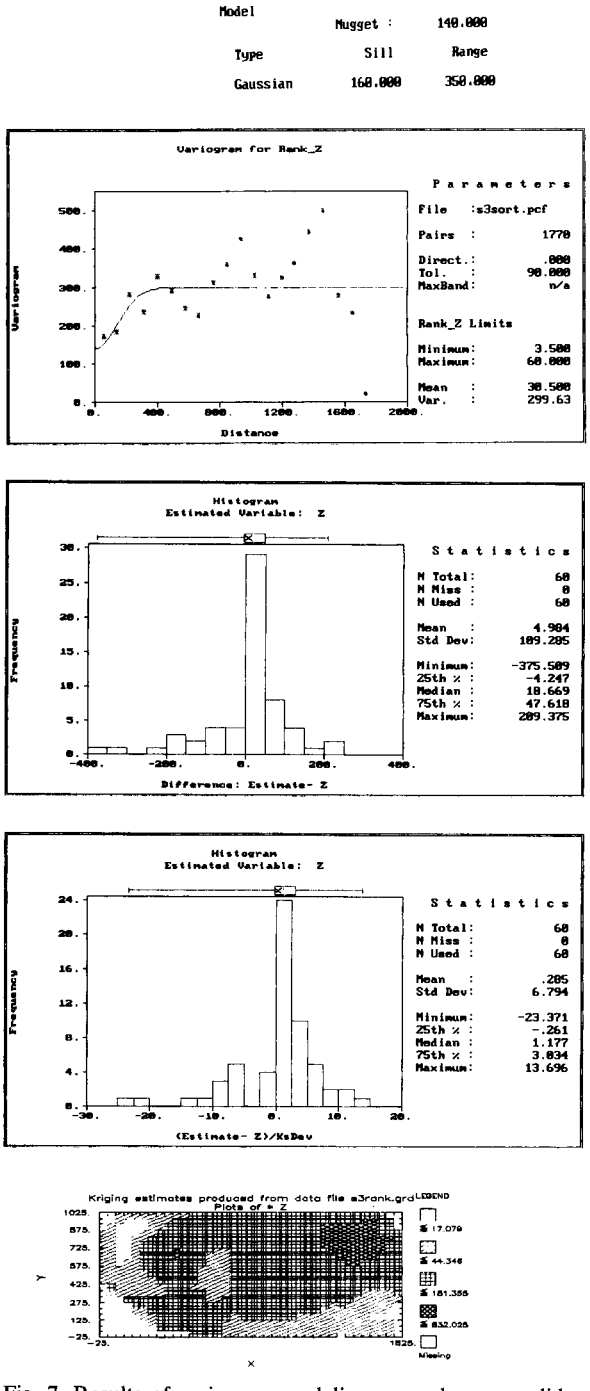


Fig. 7. Results of variogram modeling on rank, cross-validation, and Kriging on Z for the third data set from Area A.

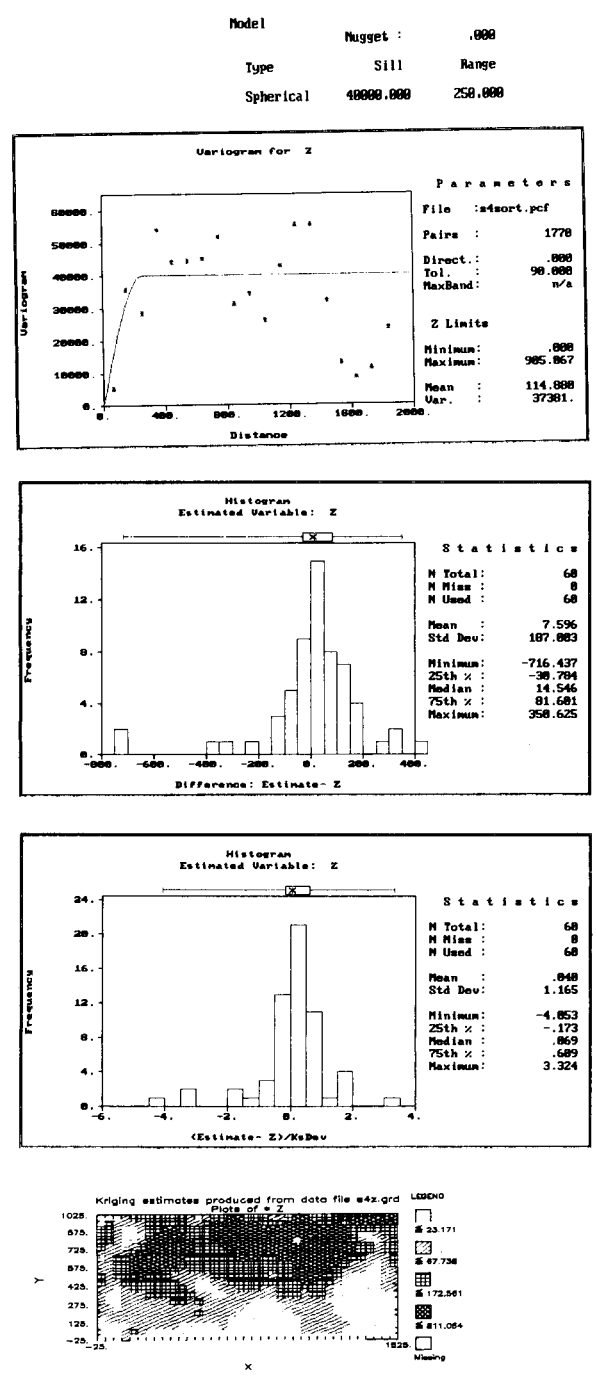


Fig. 8. Results of variogram modeling on Z, cross-validation, and Kriging on Z for the fourth data set from Area A.

conducted at the facility to develop techniques for enhanced recovery of raw materials.

During the period 1984–1989, on-site sampling investigations were conducted by the New Mexico Environmental Improvement Division (NMEID) and the United States Environmental Protection Agency (USEPA). These investigations identified lead concentrations as high as 424000 ppm in waste piles. Elevated lead levels were also found in local soil and surface sediments. Based on the criteria outlined by the Hazard Ranking System (40 CFR Chap. 1, Part 300, App. A, 7-1-87 Edition), Cal West Metals site was listed on the Nation Priorities List (NPL) on March 31, 1989.

On June 24, 1991, scientists from the Environmental Monitoring Systems Laboratory, Las Vegas (EMSL-LV) visited the Cal West site for a visual inspection, and also to collect samples for site specific calibration of X-Met 880, the field portable x-ray fluorescent spectrometer used at the Cal West Superfund site. More details about the sampling for site characterization of the site are given by Cole et al. [10].

A total of 137 samples taken on a regular grid were available for isopleth contouring of lead concentration at the site. Figure 10 shows the 137 sampling locations, and Fig. 11 shows a 3D plot of observed lead concentrations. It is seen from Fig. 11 that there is a clear trend in lead concentrations, with an increase towards the north-east portion of the site. The method of least-squares was used to fit an empirical equation to the lead concentrations as a function of (transformed) easting and northing, and then the residuals were computed.

The fitted response surface was:

$$Y_{\text{pred}} = 6.1219 + 1.0959E + 0.7665N - 0.5505E^2 - 1.0954N^2 + 0.3823EN \quad (11)$$

where $E = (\text{easting}-931.4)/418.8$; $N = (\text{northing}-681.0)/293.9$; Y_{pred} = predicted value of $\ln(Z+10)$; Z = lead concentration in ppm.

At each sampling location, the residual E was calculated by the formula:

$$E = Y_{\text{obs}} - Y_{\text{pred}} \quad (12)$$

where Y_{obs} = observed value of $\ln(Z+10)$.

The method of Ordinary Kriging is used on residuals E, and then the trend surface given by Eqn. 11 is added back to obtain estimates of $Z' = \ln(Z+10)$, and then Z is calculated from the formula:

$$Z = e^{Z'} - 10$$

We will now present the results of our analysis using the usual method of OK, and also the proposed method.

Figure 12 shows the variogram model for Z (first plot), and the cross-validation results (second and third plots). Figure 13 shows the lead concentration isopleths obtained from Kriging using this variogram model.

Figures 14 and 15 show the corresponding results using the proposed method of variogram modeling. Figures 13 and 15 show that the results of Kriging Z using the variogram on Z and the

TABLE 1

Summary statistics for standardized errors from cross-validation

	(A) Usual Ordinary Kriging: Variogram on Z, Krige Z	(B) Proposed Method: Variogram on ranks Krige Z. Values given by GEO-EAS	(C) Proposed Method: Variogram on ranks, ranks, Krige Z. Correct values = $(CF)^{0.5} \times B$
Mean	0.017	0.326	0.025
S.D.	1.055	14.249	1.079
Minimum	−3.566	−52.909	−4.008
25th %	−0.183	−2.390	−0.181
Median	0.128	1.225	0.093
75th %	0.395	5.686	0.431
Maximum	3.336	33.982	2.574

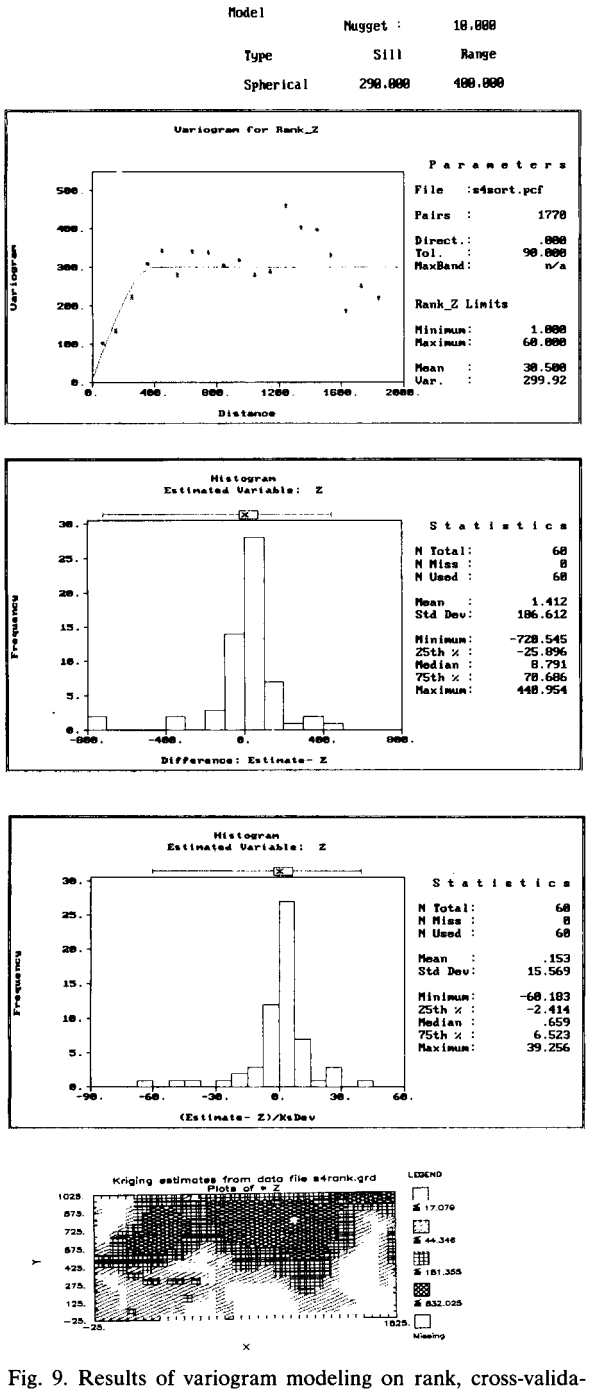
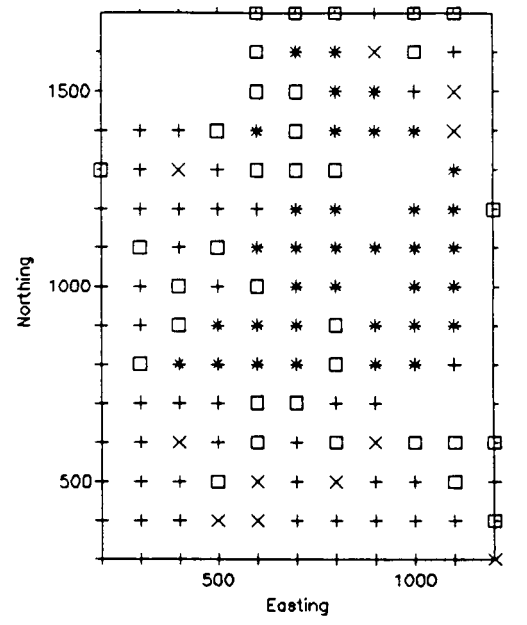


Fig. 9. Results of variogram modeling on rank, cross-validation, and Kriging on Z for the fourth data set from Area A.



1st Quartile:	.000	≦	+	≦ .000
2nd Quartile:	.000	<	×	≦ 40.000
3rd Quartile:	40.000	<	□	≦ 490.000
4th Quartile:	490.000	<	✳	≦ 42680.000

Fig. 10. Plot of sample locations at Cal West Metals NPL site (quartiles are in ppm).

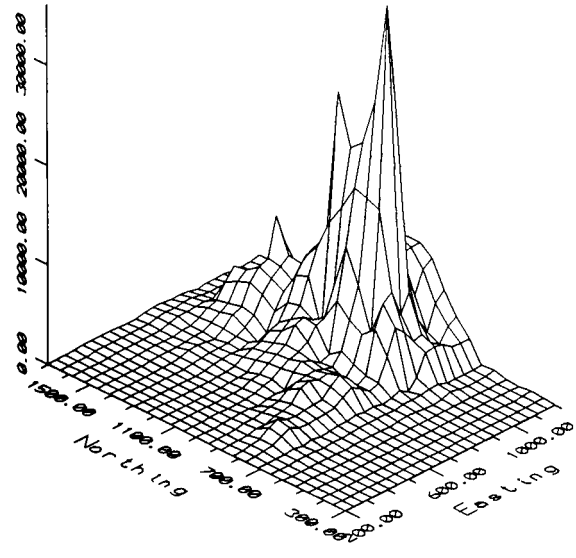


Fig. 11. Plot of observed lead concentrations (in ppm) at Cal West Metals NPL site.

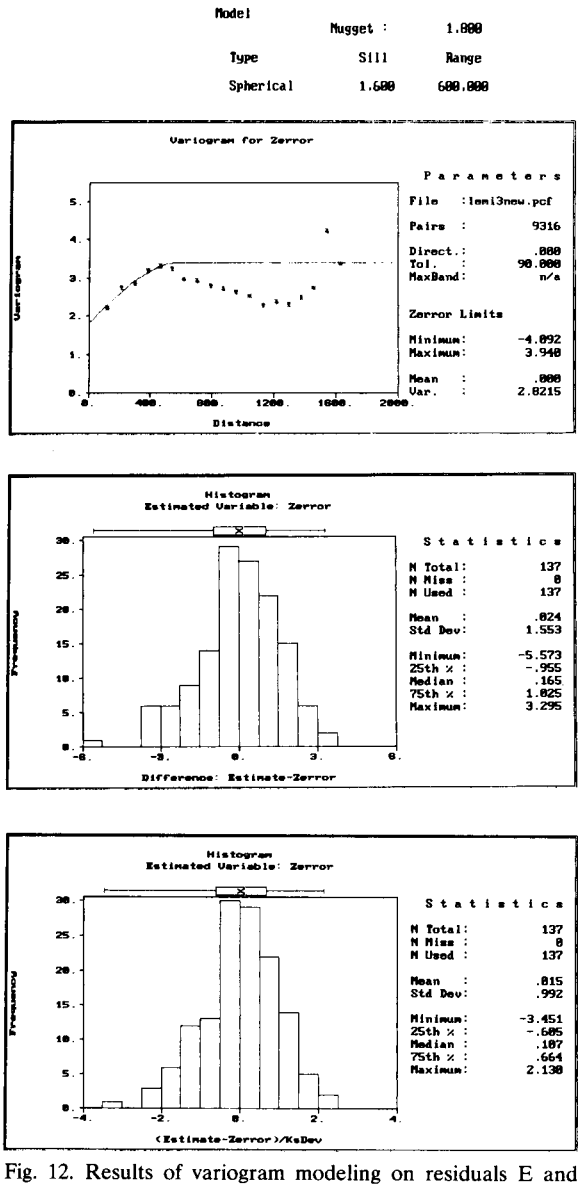


Fig. 12. Results of variogram modeling on residuals E and cross-validation for the Cal West Metals NPL site lead concentrations data.

variogram on ranks are very similar. As mentioned earlier, the variogram based on ranks is more robust than the variogram based on Z.

A.K.S. wishes to thank W.H. Cole and C.A. Kuharic of the Lockheed Engineering and Science Company, Las Vegas, and David Gonzales of the Harry Reid Center for Environmental Studies, University of Nevada, Las Vegas for their help, and to Ken Brown, EMSL-Las Vegas for his helpful comments.

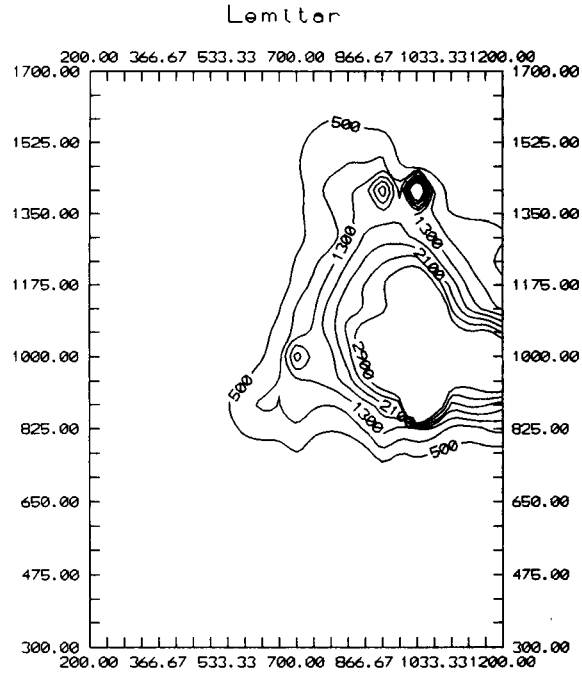


Fig. 13. Contours of lead concentrations obtained from Kriging using variogram based on lead concentrations.

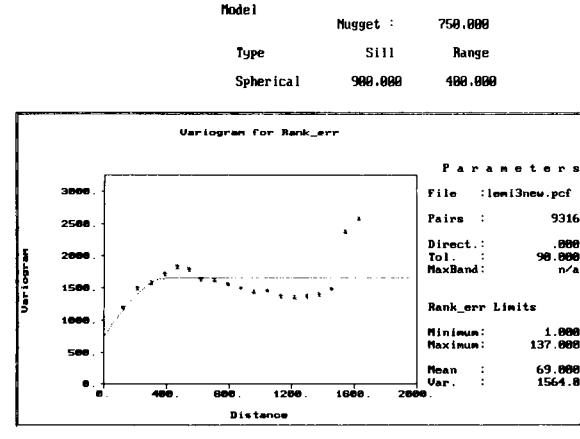


Fig. 14. Results of variogram modeling on ranks of E and cross-validation for the Cal West Metals NPL site lead concentrations data.

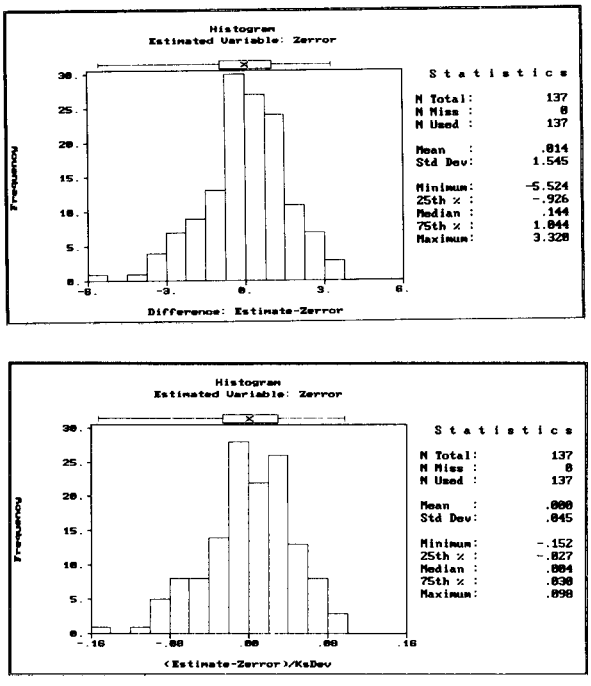


Fig. 14 (continued).

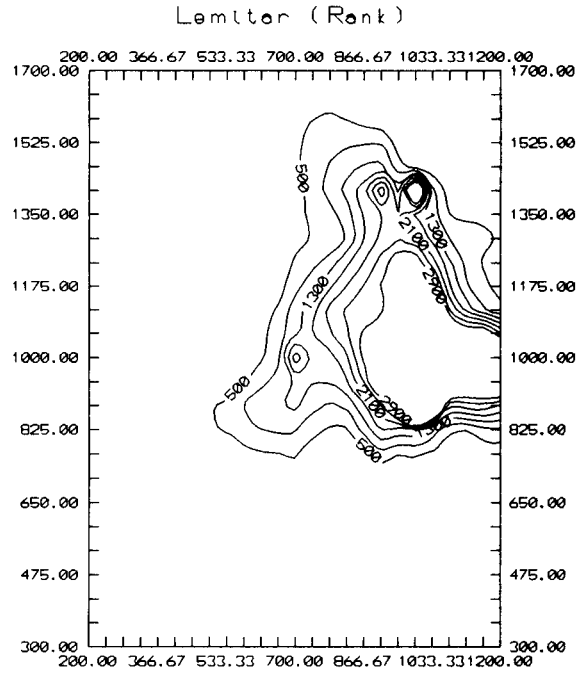


Fig. 15. Contours of lead concentrations obtained from Kriging using variogram based on ranks of lead concentrations.

REFERENCES

1 T.H. Starks, A.R. Sparks and K.W. Brown, Environmental Monitoring and Assessment, 9 (1987) 239.
2 G. Matheron, Econ. Geol., 58 (1963) 1246.
3 L.S. Gandin, Objective Analysis of Meterological Fields: GIMIZ, Leningrad, 1963 (Israel Program for Scientific Translations, Jerusalem, 1965).
4 W.J. Conover and R.L. Iman, Am. Stat., 35 (1981) 124.
5 N. Cressie and D.M. Hawkins, Math. Geol., 12 (1980) 115.
6 N. Cressie, Math. Geol., 17 (1985) 563.
7 N. Cressie, Am. Stat., 43 (1989) 197.
8 E.H. Isaaks and R.M. Srivastava, Applied Geostatistics, Oxford University Press, New York, 1989.
9 E.J. Englund, Math. Geol., 22 (1990) 417.
10 W.H. Cole, D. Gonzales, C.A. Kuharic and A.K. Singh, X-Ray Fluorescence Site Screening and Geostatistical Analysis of Soil Lead Data from the Cal West Metals NPL Site, U.S. EPA/EMSL-LV TSC-12.

Analytica Chimica Acta, 277 (1993) 267–271
Elsevier Science Publishers B.V., Amsterdam

Use of the mean quadratic error of prediction for the construction of biased linear models

Jiří Militký

Department of Textile Materials, Technical University, Liberec (Czech Republic)

Milan Meloun

Department of Analytical Chemistry, Technical University, Pardubice (Czech Republic)

(Received 30th June 1992; revised manuscript received 3rd August 1992)

Abstract

The main practical problems caused by multi-collinearity are reviewed. The biased estimators based on the generalization of principal components for avoiding multi-collinearity problems are described. The mean quadratic error of prediction criterion is used for the selection of suitable bias. Some advantages of biased regression are demonstrated on the problem of intercept estimation in a polynomial model.

Keywords: Optimization methods; Biased linear models; Mean quadratic error of prediction; Multi-collinearity

Multiple linear and non-linear chemical model building is among the most complex problems to be solved in chemometric practice. An interactive approach to model building can be divided into the following steps [1]: selection of provisional models; analysis of assumptions about the model, data and regression methods (regression diagnostics); extension and modification of the model, data and regression method; and testing the validity of the model, its predictive capability, etc.

An interactive strategy of multiple model building based on the above steps has been described [1]. Many problems in the realization of the second step are caused by strong multi-collinearity. Multi-collinearity in multiple linear regression analyses is defined as approximate linear dependences among the explanatory variables (columns of design matrix X).

Correspondence to: J. Militký, Department of Textile Materials, Technical University, Liberec (Czech Republic).

It is well known that under strong multi-collinearity the parameter estimates and hypotheses test are affected more by linear "links" between explanatory variables than by the regression model itself. The classical t-test of significance is highly inflated owing to large variances of regression coefficient estimates and the results of regression are often unacceptable.

A number of alternatives to the least-squares approach have been recommended to avoid multi-collinearity. The resulting estimators are biased, but may be preferable to classical least squares. The most popular of these are the ridge-type estimators proposed by Hoerl and Kennard [2] and several others [3].

In this paper the estimators based on generalized principal components are adopted. For suitable bias selection the criterion based on the mean quadratic error of prediction (MEP) is used. The proposed procedure of biased estimator construction is a part of the package CHEMSTAT for data analysis in chemometric practice.

SUMMARY OF LINEAR REGRESSION

The standard linear model with n observations of m explanatory variables is assumed. For an additive model of measurement errors the linear regression model has the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{1}$$

In Eqn. 1 the $n \times m$ matrix $\boldsymbol{X}$ contains the values of m explanatory (predictor) variables at each of n observations, $\boldsymbol{\alpha}$ is the $m \times 1$ vector of regression parameters and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of experimental errors; $\boldsymbol{y}$ is $n \times 1$ vector of observed values of the dependent variable.

The classical least-squares method is based on the following assumptions: regression parameters are not restricted; the regression model is linear in parameters and the additive model of measurements is valid (see Eqn. 1); design matrix $\boldsymbol{X}$ has a rank equal to n; and errors ϵ_i are independent identically distributed random variables with zero mean $E(\epsilon_i) = 0$ and diagonal covariance matrix $D(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{E}$, where $\sigma^2 < \infty$. For testing purposes it is assumed that errors ϵ_i have a normal distribution $N(0, \sigma^2)$. When these four assumptions are valid the parameter estimates $\boldsymbol{b}$ found by minimization of the least-squares criterion

$$S(\boldsymbol{b}) = \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \| \tag{2}$$

are best linear unbiased estimators (BLUE). In Eqn. 2, $\| \|$ is the symbol for Euclidean norm.

The conventional least-squares estimator $\boldsymbol{b}$ has the form

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{3}$$

The corresponding covariance is

$$D(\boldsymbol{b}) = \sigma^2 (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \tag{4}$$

From a geometrical point of view columns of design matrix $\boldsymbol{X}$ define an m-dimensional hyperplane L in n-dimensional Euclidean space E^n. The vector $\boldsymbol{X}\boldsymbol{\beta}$ and prediction vector

$$\boldsymbol{y}_{\mathrm{P}} = \boldsymbol{X}\boldsymbol{b} \tag{5}$$

lie in plane L. The prediction vector is an orthogonal projection of vector $\boldsymbol{y}$ to the plane L.

$$\boldsymbol{y}_{\mathrm{P}} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{6}$$

where $\boldsymbol{H}$ is the projection matrix. The residual vector

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{y}_{\mathrm{P}} \tag{7}$$

is orthogonal to plane L and has the minimum length. Vector $\boldsymbol{e}$ is related to projection matrix $\boldsymbol{H}$:

$$\boldsymbol{e} = (\boldsymbol{E} - \boldsymbol{H})\boldsymbol{y} \tag{8}$$

$\boldsymbol{E}$ denotes a unit matrix of order n. The variance matrix corresponding to prediction vector $\boldsymbol{y}_{\mathrm{P}}$ has the form

$$D(\boldsymbol{y}_{\mathrm{P}}) = \sigma^2 \boldsymbol{H} \tag{9}$$

and the variance matrix for residuals is

$$D(\boldsymbol{e}) = \sigma^2 (\boldsymbol{E} - \boldsymbol{H}) \tag{10}$$

Statistical analysis related to least squares is based on normality of estimates $\boldsymbol{b}$.

MULTI-COLLINEARITY

Multi-collinearity does not mean a violation of assumptions about least-squares methods. It concerns an ill-conditioning of the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ which has two consequences: the determinant of matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is near zero and some eigenvalues of matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ are near zero. This problem arises in cases when one of columns x_j of matrix $\boldsymbol{X}$ is a near linear combination of several other columns.

Multi-collinearity causes many difficulties in the inverse of matrix $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$, i.e., numerical difficulties. In addition to numerical difficulties, multi-collinearity also leads to the following statistical difficulties: non-stability of estimates b caused by the great sensitivity of parameter estimates to small changes in the data vector $\boldsymbol{y}$, the estimates $\boldsymbol{b}$ often having unacceptable signs and magnitudes, which effects their chemometric interpretation; large variances $D(b_j)$ of individual estimates cause the t-test to indicate statistical insignificance of parameters β_j; and a strong correlation between elements of estimates $\boldsymbol{b}$ means that they cannot be interpreted separately.

On the other hand, in the case of multi-collinearity the determination coefficient (square of multiple correlation coefficient) is often high and

a regression model may fit the data fairly well. For the case of data smoothing by regression models, multi-collinearity does not cause any difficulties except numerical ones. Multi-collinearity exhibits serious problems especially in regression model building (selection of explanatory variables, etc.).

Sources of multi-collinearity can be categorized into following major groups: the over-estimated regression mode; inappropriate location of experimental points; and physical constraints in model or in data.

Techniques suitable for the detection of multi-collinearity and sources of multi-collinearity have been described previously [1].

GENERALIZED PRINCIPAL COMPONENT

For convenience it is assumed that the columns of matrix $\boldsymbol{X}$ are properly scaled so that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is equal to correlation matrix $\boldsymbol{R}$. As the matrix $\boldsymbol{R}$ is symmetrical it may be expressed as a sum of eigenvalues $\tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_m$ and corresponding eigenvectors $\boldsymbol{P}_j$, $j = 1, \ldots, m$:

$$\boldsymbol{R} = \sum_{j=1}^{m} \tau_j \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}} \tag{11}$$

The inverse matrix $\boldsymbol{R}^{-1}$ can be then expressed in the form

$$\boldsymbol{R}^{-1} = \sum_{j=1}^{m} \tau_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}} \tag{12}$$

The normalized estimates $\boldsymbol{b}_{\mathrm{N}}$ can be calculated by substituting Eqn. 12 into Eqn. 3:

$$\boldsymbol{b}_{\mathrm{N}} = \sum_{j=z}^{m} \left[\tau_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}}\right] \boldsymbol{r} \tag{13}$$

The vector $\boldsymbol{r}$ is the scaled version of the vector $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ containing paired correlation coefficients between dependent and exploratory variables.

The corresponding covariance matrix has the form

$$D(\boldsymbol{b}_{\mathrm{N}}) = \sigma_{\mathrm{N}}^2 \sum_{j=z}^{m} \tau_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}} \tag{14}$$

In the case of the least squares in Eqns. 13 and 14, the constant z is taken as $z = 1$. For the principal component regression the z can be equal to 1, 2, 3, ...

From both Eqns. 13 and 14 it follows that when the eigenvalues τ_j are small the estimates $\boldsymbol{b}_{\mathrm{N}}$ and their variances are high. When some of τ_j are equal to zero the $\boldsymbol{b}_{\mathrm{N}}$ and $D(\boldsymbol{b}_{\mathrm{N}})$ are infinite. One way to avoid these difficulties is the use of generalized principal component when the small eigenvalues τ_j (or its parts) are neglected [4].

Let us denote

$$W = \sum_{j=1}^{z} \tau_j \quad \text{and} \quad E = \sum_{j=1}^{m} \tau_j$$

The criterion for leaving out the parts with too small eigenvalues then has the form

$$\mathrm{abs}(W/E) = P \tag{15}$$

where P is a selected parameter (often equal to 10^{-5}). Equality 15 cannot be generally valid for an integral z and given P. In this instance the minimum value of z for which the inequality

$$\mathrm{abs}(W/E) > P$$

is valid is selected. The summation in Eqns. 13 and 14 is then made from $z - 1$ and the term corresponding to eigenvalue τ_{z-1} is "weighted" by the factor

$$U = \left[\mathrm{abs}(W) - \mathrm{abs}(E)P\right]/\tau_z \tag{16}$$

By using this procedure, the length of estimates $\| \boldsymbol{b}_{\mathrm{N}} \|$ with their variances may be continuously decreased in dependence on increasing parameter P. Parameter P then corresponds to bias caused by neglecting some terms in Eqns. 13 and 14.

A suitable magnitude of P can be determined from the requirement for a minimum of the mean quadratic error of prediction.

SELECTION OF SUITABLE P

One of the main properties of regression models is the good predictive ability. This predictive ability can be adopted also for selection of, in some sense optimum, parameter P.

TABLE 1

Experimental data

x	25	35	45	55	65	75	85	95	105	115
y	150	160	170	190	210	230	270	310	370	450

The predictive ability in a linear regression model can be characterized by the mean quadratic error of prediction (MEP), defined generally by

$$\mathrm{MEP} = \sum_{i=1}^{n} \left[y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{b}(i) \right]^2 / n \qquad (17)$$

where $\boldsymbol{b}_{(i)}$ is the estimate of regression model parameters when all points except the ith (ith row $\boldsymbol{x}_i$ of matrix $\boldsymbol{X}$) are used. The statistics MEP uses a prediction $y_{Pi} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{b}_{(i)}$ which was constructed without information about the ith point.

The estimate $\boldsymbol{b}_{(i)}$ can be calculated from the least-squares estimate $\boldsymbol{b}$:

$$\boldsymbol{b}_{(i)} = \boldsymbol{b} - \left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \boldsymbol{x}_i e_i \right] / [1 - H_{ii}] \qquad (18)$$

where H_{ii} is the diagonal element of the projection matrix $\boldsymbol{H}$.

After substitution from Eqn. 18 into Eqn. 17, the following simple relation results:

$$\mathrm{MEP} = n^{-1} \sum_{i=1}^{n} e_i^2 / (1 - H_{ii})^2 \qquad (19)$$

For a selected P it is possible to calculate values of H_{ii} from Eqn. 6 and then the MEP criterion from Eqn. 19.

A suitable P corresponds to some minimum value of MEP. For the selection of this value of P a very simple strategy can be used: for $P \approx 10^{-30}$ the MEP_1 is calculated; for $z = 2, 3, \ldots$ the MEP_z are calculated until $\mathrm{MEP}_z < \mathrm{MEP}_{z-1}$; and in the interval $W_{z-1}/E \leqslant P \leqslant W_z/E$ the optimum P is selected by the interval halving method.

This procedure is very simple and requires only one decomposition of matrix $\boldsymbol{R}$. The calculated P do not correspond generally to a global minimum but parameter estimates and the statistical characteristics are greatly improved.

In the program package CHEMSTAT (TriloByte) the generalized principal component is used and the MEP criterion is computed. Then the trial-and-error procedure can be adopted for selecting a suitable P.

EXAMPLE

Many problems in chemometrics concern an approximation of instrumental data of convex (or concave) increasing (or decreasing) values by a polynomial so that this polynomial fulfils the condition of remaining the shape of the data. For solution of these types of problems the generalized principal component with optimum P minimizing the MEP can be used.

In connection with modelling a mechanical problem the intercept term was important. The

TABLE 2

Results for principal component regression

z	τ_z	P	MEP	b_7	s_7
1	9.30×10^{-10}	1.5×10^{-10}	380.20	195.5	355.9
2	4.53×10^{-7}	7.6×10^{-8}	15.60	152.3	81.2
3	7.38×10^{-5}	1.2×10^{-5}	9.12	138.7	27.6
4	5.65×10^{-3}	9.5×10^{-4}	8.85	128.3	8.7
5	2.35×10^{-1}	4.0×10^{-2}	7.63	131.2	3.8
6	5.76	1.0	10.29	136.1	2.4

TABLE 3

Regression results for least-squares (LS) and generalized principal component (GPC) regression

Method	P	MEP	D [a]	b_7	s_7
LS	10^{-30}	380.30	0.99960	195.5	355.9
GPC	0.28	6.436	0.99953	132.4	3.48

[a] Square of multiple correlation coefficient.

data are strictly convex (see Table 1) and the regression model was specified as a polynomial of the sixth degree (based on some formal and theoretical assumptions) [5]:

$$E(y) = \sum_{j=1}^{6} b_j x^j + b_7$$

The parameter b_7 is equal to the intercept. Table 2 gives results for the principal component regression.

Table 3 gives estimates b_7, standard deviations s_7 and determination coefficient D found by the classical least-squares procedure ($P = 10^{-30}$) and generalized principal components $P = 0.28$ for which the MEP criterion was the smallest.

From Tables 1 and 3 it is obvious that the intercepts from LS do not correspond to the experimental data. The estimate b_7 is higher than the values y_1, y_2, y_3 and y_4, which indicates that the proposed model has some minimum between the origin and point (x_1, y_1). The corresponding standard deviation s_7 is very high so that the estimate b_7 is very imprecise. The parameter b_7 calculated by generalized principal components is acceptable and precise.

Conclusion

The method of generalized principal components in combination with the MEP criterion is very attractive for constructing biased models. It can be also used for achieving such estimates which keep the model course corresponding to the data trend especially in polynomial-type models. This method is implemented in the software package CHEMSTAT (TriloByte) [6].

REFERENCES

1 M. Meloun, J. Militký and M. Forina, Chemometrics 2, Interactive Model Building and Testing on IBM PC, Horwood, Chichester, 1992.
2 A.E. Hoerl and P.W. Kennard, Technometrics, 12 (1970) 55.
3 R.R. Hocking, F.M. Speed and M.J. Lynn, Technometrics, 18 (1976) 425.
4 D.M. Marquardt, Technometrics, 12 (1970) 591.
5 K. Květoň, unpublished report, Czech Technical University, Prague, 1988.
6 CHEMSTAT Version 2.0, TriloByte, Pardubice, Czechoslovakia, 1991.

Analytica Chimica Acta, 277 (1993) 273–287
Elsevier Science Publishers B.V., Amsterdam

Data processing using neural networks

T.B. Blank and S.D. Brown

Department of Chemistry and Biochemistry, Brown Laboratories, University of Delaware, Newark, DE 19716 (USA)

(Received 28th September 1992)

Abstract

Layered, feed-forward neural networks are powerful tools particularly suited to the analysis of non-linear, multivariate data. In this paper, some problems in multivariate calibration and pattern recognition are used to compare the performance of these neural networks, as trained by back-propagation, with some more well-established chemometric techniques. Simulated data sets are used to probe the capabilities of neural networks in applications where the chemometric methods are known to excel, and to fail. Results obtained from these studies support the conclusion that results obtained from application of layered, feed-forward networks are competitive with those obtained from traditional chemometric methods.

Keywords: Multivariate calibration; Pattern recognition; SIMCA method; UNEQ method

Over the last several years an increasing number of studies have examined artificial neural networks for solving modeling problems in analytical chemistry. Several efforts in multivariate calibration using mixture spectra and concentration data have been reported. Initially, linear, multivariate calibration relations were studied [1], but more recently spectra systems containing non-linearities due to stray light [2] and matrix effects [3,4] have received attention. Artificial neural networks have also been shown to be effective pattern classifiers. The classification of different, linearly separable algae species was performed using an artificial neural network on data from a flow cytometer [5]. A neural network has also been used to determine the presence or absence of functional groups in a large number of organic compounds from their infrared spectra [6].

The strength of modeling with artificial, layered, feed-forward artificial neural networks lies in the flexibility of the distributed soft model defined by the weights of the network. Both linear and non-linear functions may be defined by suitably configuring the neural network. The addition of a hidden layer with appropriate transfer functions converts a simple, two-layer (input and output) linear neural network to a three-layer network capable of very general modeling. It has been demonstrated that three-layer feed-forward neural networks with sigmoid transfer functions are capable uniformly approximating *any* continuous function defined on the d-dimensional space [7]. The fitting of non-linear functions in particular is often accomplished through creating superpositions of these scaled, rotated, and shifted sigmoidal functions. Linear models can be developed by using linear transfer functions to connect layers. The application of artificial neural networks in linear problems has received little attention because of the availability of comparatively less complex methods, including partial least squares regression (PLS), and principal components regression (PCR) in multivariate calibration, and linear discriminant analysis in pattern

Correspondence to: S.D. Brown, Department of Chemistry and Biochemistry, Brown Laboratories, University of Delaware, Newark, DE 19716 (USA).

classification, all of which have repeatedly been shown to be very effective in dealing with linear systems.

Neural networks are not the only method for modeling multivariate, non-linear relationships, however. Several chemometric methods are also available. Quadratic PCR [8] and polynomial PLS [9], modifications of the more familiar linear PCR and PLS, make use of a polynomial relationship to relate the independent response and dependent concentration variables in non-linear, multivariate calibration. While the non-linear PCR and PLS methods share the computational simplicity of the linear PCR and PLS methods, a possible weakness of the non-linear modeling performed by these soft modeling methods lies in the use of a simple polynomial expansion in the fitting of complex, non-linear response surfaces. Non-linear relationships also occur in classification problems. Distributions of patterns which are distinguished by non-linear class boundaries occur in many data sets; one common situation where non-linear class boundaries occur is the class-in-class problem known as the "asymmetric case", where one class is entirely contained within a second, more diffusely distributed class. Two pattern classification methods, soft-independent modelling of class analogy (SIMCA) [10], and the UNEQ classifier [11], are often used to determine class boundaries for class-in-class data by creating class models from statistical distribution information. In contrast to these methods, neural networks are not dependent on specific non-linear functional forms or on assumptions about the statistical distribution of samples.

This paper will provide an overview of some research that demonstrates the advantages of neural networks in comparison with traditional chemometric methods in the solution of some problems in multivariate calibration [12] and classification of class-in-class data [13]. The first part of the paper will consider the use of neural networks in non-linear calibration. Three simulated non-linear multivariate calibration data sets with different functional relationships between the dependent and independent variables will be used to investigate the relative performance of neural networks, polynomial PLS and quadratic PCR on unknown multivariate relationships of differing degrees of non-linearity and in different amounts of additive noise. Absorption spectra taken on aqueous mixtures or rhodamine 6G and rhodamine B will also be used to compare neural networks, polynomial PLS, and quadratic PCR for non-linear, multivariate correction of strong fluorescence of analytes. The second part of the paper concerns classification of data separated by non-linear class boundaries. Simulated class-in-class data sets will be used to compare SIMCA, UNEQ, and neural networks in the classification of objects inside and outside the 95% confidence spheroid of homogeneously distributed, trivariate, N(0,1) data. In addition, the same methods will also be compared in the case where the training data inside the spheroid are non-homogeneously distributed.

THEORY

Back-propagation

Back-propagation of error is an effective method for training the weights of layered, feed-forward neural networks which contain hidden layers [14]. The theory of back-propagation is covered briefly here to put an improved training method in context. More complete descriptions, along with some useful discussions, have been provided in an overview [15], and a review [16]. The back-propagation of error paradigm is based on the assumption that all weights contribute to some portion of the output error, and that weight corrections should be proportional to the output error contributed by each weight. In the first back-propagation step, the assigned error associated with the output layer weights connected to output node j is obtained by scaling the output error with the value of the derivative of the transfer function at the sum of imputs, x_j, to output node j

$$\delta_j = (t_j - a_j)\mathrm{f}'(x_j) \tag{1}$$

In Eqn. 1, t_j is the target value at output node j, and a_j is the actual output, or "activation", at the same node. In the second back-propagation step, the errors associated with hidden layer weights

connected to hidden layer node i are determined from scaling the weighted sum of the scaled output errors with the value of the derivative of the transfer function at the sum of inputs, x_i, to the hidden layer node i

$$\delta_i = \left[\sum_{j=1} \delta_j w_{ji}\right] \mathrm{f}'(x_i) \tag{2}$$

In the third back-propagation step, weight updates are made by using the scaled errors and the Widrow–Hoff delta rule

$$\Delta w_{ji} = \eta \delta_j a_i \tag{3}$$

The change in each connection weight between hidden layer node i and output layer node j is determined by a universal learning coefficient, η, the activation a_i of node i in the hidden layer, and δ_j, the scaled error associated with weights connected to node j in the output layer. Corrections to the hidden layer weights are made analogously using the hidden layer error terms and activations from the connected input layer nodes.

Finding the weights in a back-propagation neural network requires an optimization of the error cost function for a given pattern p over all m output nodes

$$E_\mathrm{p} = \tfrac{1}{2} \sum_{j=1}^{m} (t_j - a_j)^2 \tag{4}$$

Given the logistic sigmoid transfer function, the output at node j, a_j, is computed using the sum of weighted inputs to node j, $x_j = \Sigma a_i w_{ji}$

$$a_j = \mathrm{f}(x_j) = \frac{1}{1 + \exp(-x_j)} \tag{5}$$

The scaled error associated with node j in the output layer is

$$\delta_j = (t_j - a_j)\mathrm{f}'(x_j) = -\left(\frac{\partial E_\mathrm{p}}{\partial a_j}\right)\left(\frac{\partial a_j}{\partial x_j}\right) \tag{6}$$

The activation of node i in the hidden layer can also be expressed as

$$a_i = \frac{\partial x_j}{\partial w_{ji}} \tag{7}$$

Substitution of Eqns. 6 and 7 into Eqn. 3 gives

$$\Delta w_{ji} = -\eta\left(\frac{\partial E_\mathrm{p}}{\partial a_j}\right)\left(\frac{\partial a_j}{\partial x_j}\right)\left(\frac{\partial x_j}{\partial w_{ji}}\right) = -\eta \frac{\partial E_\mathrm{p}}{\partial w_{ji}} \tag{8}$$

which corresponds to gradient descent in n-dimensional weight space with a fixed universal learning coefficient η. Two successive applications of the chain rule defined in Eqns. 6 and 7 yield the same result for hidden layer weight corrections.

One of the main problems with back-propagation neural networks is that long training sessions are often required in order to find an acceptable weight solution because of the well-known difficulties inherent in gradient descent optimization. Two-dimensional error surface representations can be used to illustrate the conditions which produce slow learning in back-propagation. Slow learning occurs when the error surface encountered has a small slope as shown in Fig. 1. Note how the optimization step size is repeatedly attenuated by successively decreasing gradient magnitudes. On this kind of error surface, the back-propagation learning coefficient should be increased to compensate for the decreasing magnitude of the gradient, but a smaller learning coefficient may be essential for convergence when steps are being taken in the vicinity of the minimum as shown in Fig. 2. An empirically-derived time-scheduling of learning coefficients can improve the convergence of back-propagation training of a neural network, but unless the desired minimum is symmetric with respect to the weight axes, a scheduled "optimal" universal

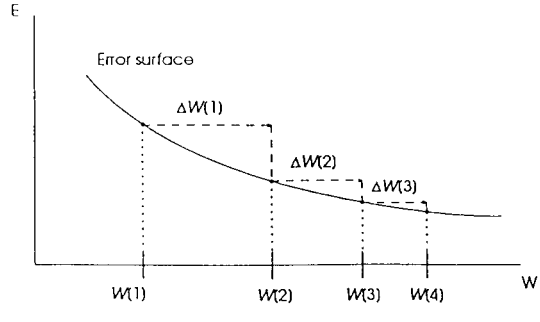


Fig. 1. Example of a 2-dimensional error surface region where large back-propagation learning coefficients are beneficial.

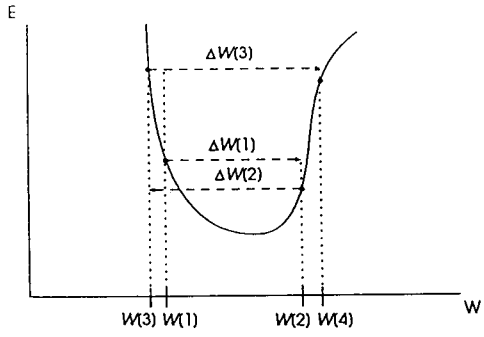


Fig. 2. Example of 2-dimensional error surface region which requires a small back-propagation learning coefficient for reliable convergence to the minimum.

learning coefficient will be, at best, only a compromise between the best coefficients for each weight axes. A helpful modification which helps to resolve this dilemma involves the use of a momentum term α which adds a fraction of the previous weight correction to the gradient, generating the new weight correction.

$$\Delta w_{ji}(k) = -\eta \frac{\partial E_p}{\partial w_{ji}} + \alpha\, \Delta w_{ji}(k-1) \qquad (9)$$

Including momentum in weight updates has been shown to contribute to more rapid convergence of the neural network by increasing the effective learning coefficient by a factor of $1/(1-\alpha)$ in regions of relatively constant gradient, and by damping the oscillatory behavior of weight updates at the bottom of steep valleys, where gradient components alternate in sign [17].

Extended delta-bar delta back-propagation

A substantial reduction in training times of multi-layer, feed-forward neural networks applied to several benchmark problems can be achieved by using individual, self-adjusting, learning and momentum coefficients for each weight axis in the back-propagation [18]. The use of weight-specific, self-adjusting learning and momentum coefficients results in faster training, and more robust convergence properties than training the network with back-propagation corrected with momentum [19]. This modification to back-propagation training uses the same delta weight update rule as back-propagation with momentum, except that individual, self-adjusting learning and momentum coefficients are used in updating each connection weight

$$\Delta w_{ji}(k) = -\eta_{ji} \frac{\partial E}{\partial w_{ji}} + \alpha_{ji}\, \Delta w_{ji}(k-1) \qquad (10)$$

To improve the temporal stability of the learning process, the weight updates indicated above are accumulated over the entire training set or "epoch". The adjustment of learning and momentum coefficients is based on the sign of the product function delta-bar delta

$$\bar{\delta}(k-1)\delta(k)$$

where $\bar{\delta}$, the weighted average gradient component, is obtained using a convex weighting factor, Θ, and the two previous gradient components

$$\bar{\delta}(k-1) = (1-\Theta)[\delta(k-1) - \Theta\delta(k-2)] \qquad (11)$$

as described in the Appendix. When the product indicated by $\bar{\delta}(k-1)\delta(k)$ is positive, the individual learning and momentum coefficient increases are exponentially decreasing functions of the magnitude of the gradient component along each weight axis. Alternately, the learning and momentum coefficients are reduced by a multiplicative factor, ϕ, when $\bar{\delta}(k-1)\delta(k)$ is negative. The result is slow growth, or even attenuation, of the coefficients around minima, and fast growth in regions of the error surface with a small slope of constant sign. These modifications accelerate back-propagation training by increasing learning rates on flat regions of the error surface and by stabilizing learning around narrow minima.

EXPERIMENTAL

Software

The extended delta-bar delta back-propagation training routines contained in the NeuralWorks Professional II software package were used for developing artificial neural networks throughout this research. The artificial neural networks created by the NeuralWorks package were run on the Apple Macintosh IIfx computer, where some 600 000 presentations could be obtained in one

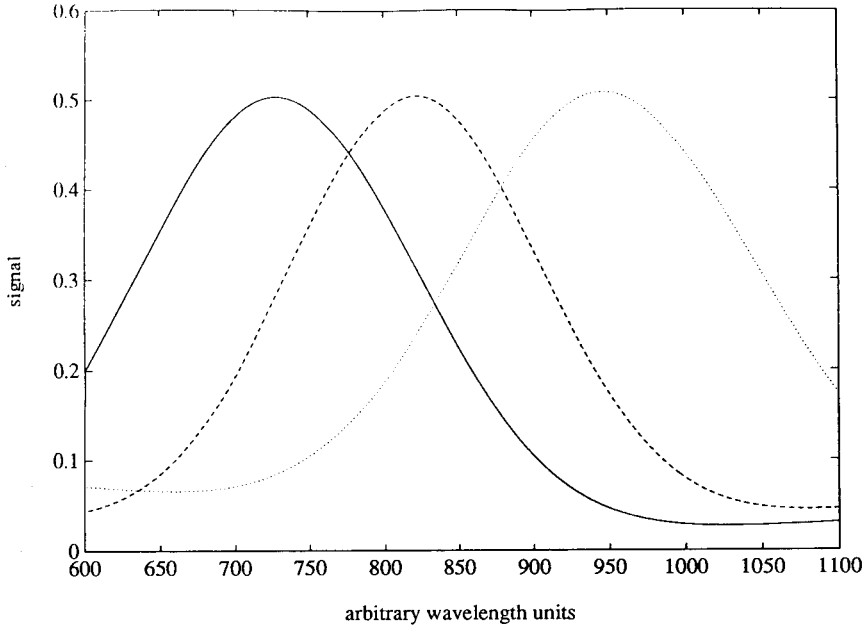


Fig. 3. Gaussian peak models used in simulated multivariate calibration data.

hour. Training was limited to approximately 10 h for each attempt. Care was taken to avoid over-training of the networks.

Matrix multivariate calibration methods, including linear and non-linear PLS and PCR, and the UNEQ and SIMCA pattern classifiers were

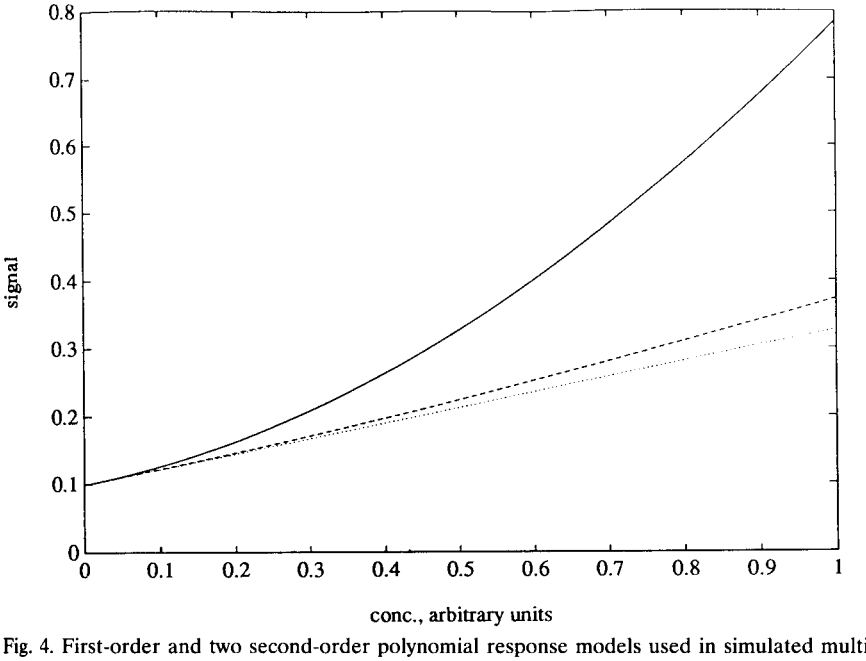


Fig. 4. First-order and two second-order polynomial response models used in simulated multivariate calibration data.

written in the MATLAB programming language following published algorithms [8–11]. These were also run on an Apple Macintosh IIfx.

Multivariate calibration data

Three three-component multivariate synthetic data sets were generated using one linear, and two different second-order polynomial relationships between the response and concentration variables. The Gaussian peak models of the three analytes in Fig. 3 were used along with 150 uniformly distributed concentration patterns, in which each of the three component concentrations ranged from 0.01 to 1. Individual component spectra were produced from the product of Gaussian peak vectors and polynomial concentration functions of the form

$$\text{component spectra} = s(i)\left[a_i C_{ij}^2 + b_i C_{ij} + \text{int}(i)\right] \quad (12)$$

In Eqn. 12, the C_{ij} are the concentrations of component i, sample number j, and a_i, b_i and int(i) are the coefficients of the polynomial model for component i, while $s(i)$ is the spectrum of component i at unit concentration. Individual component spectra for each sample were summed in order to obtain 150 three-component mixture spectra. The strong second-order model coefficients are given in Table 1. The mild second-order model coefficients were obtained by multiplying the strong second-order quadratic terms (a_i) by one tenth. The linear model used in this study was created from the first-order (b_i) and intercept terms of the strong second-order model. In Fig. 4, the responses of the strong and mild second-order polynomial models using coefficients from component number one are plotted along with the linear response model.

TABLE 1

Polynomial model coefficients: strong second-order model

Component 1	Component 2	Component 3	Term [a]
0.4583	0.9444	0.9039	a (second-order)
0.2282	0.6983	0.6251	b (first-order)
0.0986	0.0064	0.0940	intercept

[a] Term in Eqn. 12.

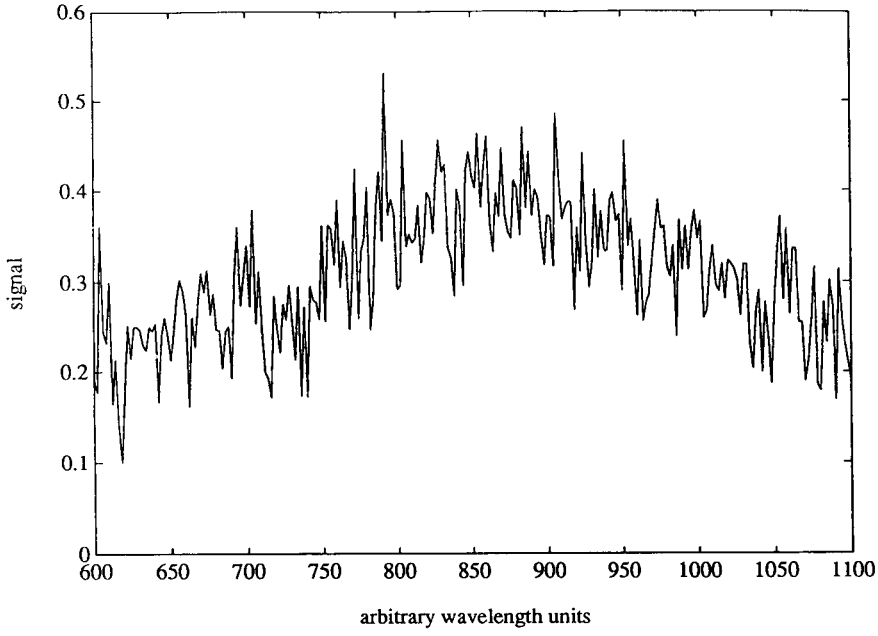


Fig. 5. High noise on a spectrum with relatively low signal strength.

Three levels of constant-magnitude, zero-mean, Gaussian white noise were also added to the data in order to study the effect of noise levels on the accuracy of prediction of the different methods. Low levels of noise, medium levels of noise, and high noise levels corresponded to mean noise levels at 0.32, 1.6 and 3.2% relative to the maximum signal in each data set, respectively. Figure 5 is an example of the degree of noise to be expected at low signal strengths with high noise.

An exponentially non-linear, two-component data set was also generated in order to study the capability of the different calibration methods in fitting functions with changing curvature. This two-component system used generalized fluorescence response models with non-linear behavior. The simulated fluorescence response model was

$$\text{intensity}(i, j) = \{1 - 10^{-[\text{const}(j)\cdot\text{conc}(i)]}\} \qquad (13)$$

The exponential constants, simulating a product of the molar absorptivity and path length of the spectrometer cell, were chosen as 1.2 and 1.7 for components one and two, respectively. The combination of the exponential constants, and the concentration range of 0.01–1, form the highly non-linear response models which are plotted along with the strong second-order polynomial response model in Fig. 6. As in the case of the polynomial data sets, these response models were used along with the Gaussian peak models, and the uniformly distributed concentrations of the first two components from the preceding polynomial data set to generate 150 simulated mixture spectra.

A two-component system of rhodamine 6G and rhodamine B mixtures, showing non-linear concentration–response relations due in part to overlap between the absorption and fluorescence emission spectra, was also examined. Eighty mixtures of laser dye-quality rhodamines with uniformly distributed concentrations between 0.5 and 20 mg l^{-1} were made by weighing aliquots of 40 mg l^{-1} stock solutions and the water diluant to four significant figures. Absorption spectra between 510 and 560 nm, collected with a Hewlett Packard 8452A spectrometer, were used in the calibration study.

Data pretreatment and method validation used in the calibration studies are described in Ref. 12.

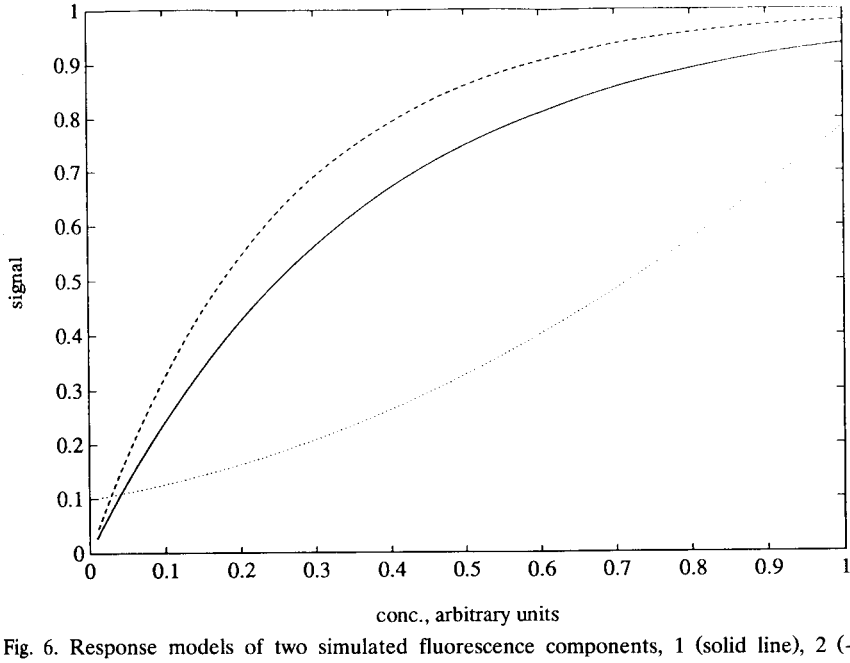


Fig. 6. Response models of two simulated fluorescence components, 1 (solid line), 2 (— — —), vs. the strong second-order polynomial response model (· · · · · ·).

Pattern classification data

The simulated data used to study the homogeneous asymmetric case was randomly generated, trivariate, N(0,1) data. Class members were identified as those data falling inside a sphere with a radius of 2.794, and a centroid of (0,0,0). This corresponds to the 95% confidence surface of trivariate, N(0,1) data. A larger set of trivariate, N(0,1) data was used to obtain an adequate sampling of outliers. Data falling outside the sphere were identified and saved for the outlier training and test set data. Because the outliers were also obtained from the distribution tails of trivariate N(0,1) data, the highest density of outliers occurred at the decision boundary, allowing a highly sensitive comparison of the sensitivities of the various methods to outliers.

A non-homogeneous training set was also created by adding a mean offset to two thirds of another set of randomly generated, trivariate, N(0,1) data. Class data were still required to fall inside the class sphere defined in the homogeneous data set, but the distribution of class members was not centered in the sphere.

Real data used in pattern classification was the well-known iris data set [20]. Three different species of iris flowers were classified using four distance metric inputs corresponding to sepal and petal lengths and widths. Seventy-five training and seventy-five test set pattern pairs were equally distributed among the three linearly-separable flower classes.

RESULTS AND DISCUSSION

Neural network design

The effective application of neural networks in multivariate calibration requires strategies similar to those used with other chemometric calibration techniques. In order to obtain a general solution, the problem must be over-determined in an object/feature sense. While features or variables determine the number of degrees of freedom in matrix regression methods, the degrees of freedom in a neural net are related to the number of connections, and to the number of hidden nodes [21]. The proofs given in Ref. 21 are quantitatively valid only for classifications using hard limiter transfer functions, but a qualitative general rule for calibration mappings using continuous sigmoidal transfer functions is to minimize the number of connections and hidden nodes while using a substantial number of training samples. These goals were accomplished by reduction of the input variables, and by an empirical determination of the minimum number of hidden nodes required to obtain best training performance. Overall, the number of adjustable, weighted connections used in the neural models was held to a maximum of two-thirds of the number of training samples presented to the network.

Reduction in the number of input variables can greatly reduce the overall number of connections in spectroscopic multivariate calibration. A reduction of inputs was accomplished using a principal component extraction of scores from the mixture spectra. A reduction of 251 to 2 or 3 inputs was realized using synthetic multivariate calibration data sets, and 26 inputs were reduced to 4 in the multivariate calibration study of rhodamine mixtures.

Determination of the number of hidden nodes was based on a comparison of training rates during the period just after the initial rapid learning phase. Hidden nodes were added until learning rates failed to show a significant improvement during this phase. The number of hidden nodes used in the final network model was then decremented by one.

The neural network configurations used in this research follow the identification code described as follows: An n digit number was used in which each digit corresponds to the number of nodes in the succeeding n layers of the network. For example, a 341 network had 3 input nodes, 4 hidden nodes, and 1 output node. The series of letters following the number were used to designate the transfer function used in the hidden and output layers, respectively. For example, a 2432ssl network contained 2 input nodes, 4 nodes in the first hidden layer with sigmoid transfer functions, 3 nodes in the second hidden layer with sigmoid transfer functions, and 2 nodes in the output layer with linear transfer functions.

Multivariate calibration

One hundred and fifty spectral mixture–concentration pattern pairs were split into one hundred training pairs and fifty test pairs for the four different data sets. The large number of training and test patterns allowed for a thorough study of the performance of the different calibration methods. The percent standard error of prediction (%SEP) was used to evaluate the prediction performance of the various multivariate calibration methods

$$\%\mathrm{SEP}_j = \frac{100}{\bar{c}_j}\left[\frac{\sum_{i=1}^{n}\left(\hat{c}_{ij} - c_{ij}\right)^2}{p}\right]^{1/2} \qquad (14)$$

Eqn. 14 gives the %SEP for the jth component where p is the number of test concentration mixtures, $\hat{c}_{ij}$ and c_{ij} are the predicted and actual concentrations of component j in sample mixture i, and $\bar{c}_j$ is the concentration mean of component j in the test set. In all synthetic data studies, the %SEP values given reflect prediction errors calculated over all analyte components using the concentration mean of the entire test set.

The results from the linear data set shown in Table 2 indicate that PLS, PCR, and the 313ll neural network were essentially equivalent in producing adequate calibration relations for this data set. The results obtained using the mild second order data are listed in Table 3. Quadratic PCR outperformed polynomial PLS and the 363sl network on noise-free data. While the 363sl network could be trained longer to produce lower prediction errors than quadratic PCR if desired, neural network results reported here were obtained within a ten-hour training limit (approximately 6×10^6 presentations). Better predictions obtained with quadratic PCR as compared with polynomial PLS appeared to be due to the inclusion of score cross-terms in the augmented PCR scores matrix. This conclusion was supported by the increase in SEP values of quadratic PCR when the score cross-terms were removed. In effect, the quadratic PCR method used a full second-order polynomial score matrix with cross-terms, and polynomial PLS used a quadratic inner relation without cross-terms. Not surprisingly, added noise reduced the range in the prediction errors of the different methods. In Table 4 it can be seen that a substantial change in relative prediction errors resulted when the three calibration methods were applied to the strong second-order data. At this level of non-linearity, the 363ss

TABLE 2

Calibration of the linear data set

Noise level	Error (%SEP)		
	313ll NN	Linear PLS	Linear PCR
Noise-free data	4.02×10^{-5}	1.35×10^{-2}	6.42×10^{-7}
Low noise (0.32% of max. signal)	0.475	0.477	0.474
Medium noise (1.6% of max. signal)	2.81	2.86	2.79
High noise (3.2% of max. signal)	5.73	5.91	5.61

TABLE 3

Calibration of the mild second-order data set

Noise level	Error (%SEP)		
	363sl NN	Polynomial PLS	Quadratic PCR
Noise-free data	0.261	1.55	0.154
Low noise (0.32% of max. signal)	0.577	1.61	0.503
Medium noise (1.6% of max. signal)	2.77	3.09	2.78
High noise (3.2% of max. signal)	4.54	5.04	4.49

TABLE 4

Calibration of the strong second-order data set

Noise level	Error (%SEP)		
	363ss NN	Polynomial PLS	Quadratic PCR
Noise-free data	0.586	7.07	2.60
Low noise (0.32% of max. signal)	0.850	7.08	2.64
Medium noise (1.6% of max. signal)	2.94	7.41	3.49
High noise (3.2% of max. signal)	5.17	9.11	5.35

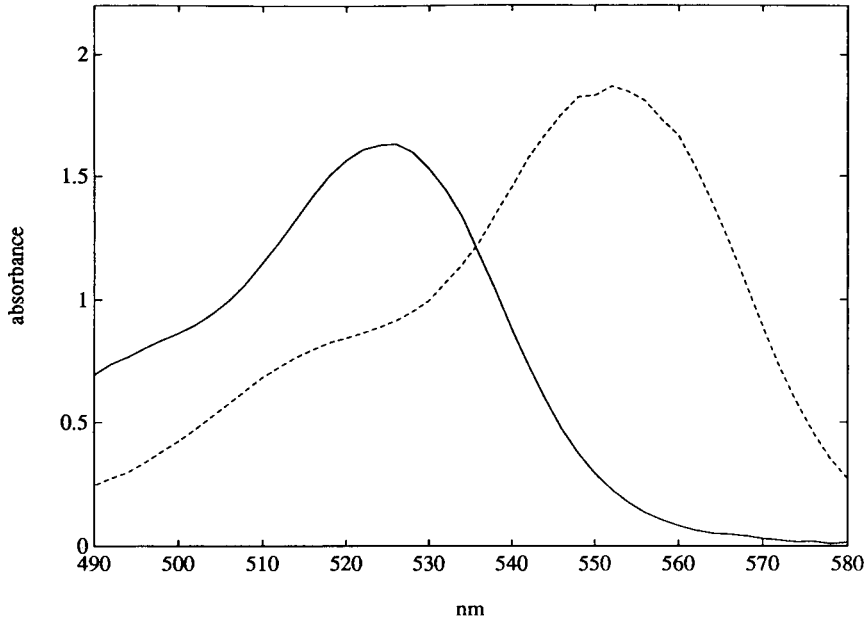


Fig. 7. UV–visible spectra of rhodamine 6G (solid line), and rhodamine B (— — —).

neural network outperformed quadratic PCR by a factor of 4, and polynomial PLS by a factor of 12. For this data set, though, the addition of medium or high noise to the data reduced the performance of the neural network to that of quadratic PCR. The results for the most strongly

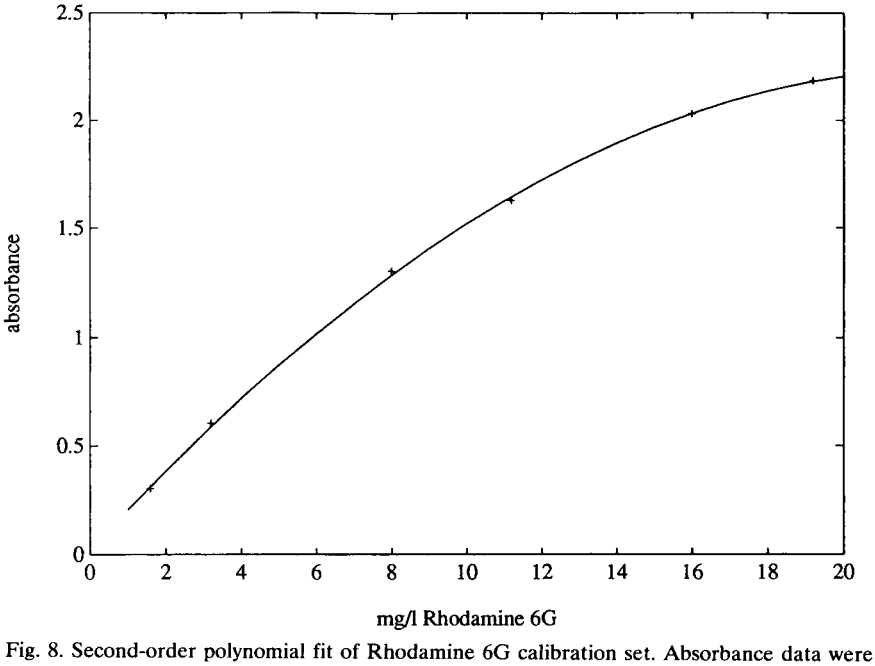


Fig. 8. Second-order polynomial fit of Rhodamine 6G calibration set. Absorbance data were taken at 530 nm.

TABLE 5

Calibration of simulated fluorescence data set

Noise level	Error (%SEP)		
	2432ssl NN	Polynomial PLS	Quadratic PCR
Noise-free data	0.571	18.55	10.14

non-linear data set, the two-component simulated fluorescence data, are given in Table 5. The 2432ss neural network, with two hidden layers, outperformed both regression methods by a wide margin. This result is probably due to the high degree of non-linearity of the response function, and to the inability of a second-order polynomial to fit a function of rapidly changing curvature.

Multivariate calibration was also performed for the rhodamine mixtures by using neural networks as well as linear and non-linear PCR and PLS. Aqueous solutions of rhodamine B and rhodamine 6G have overlapped UV–visible spectra which are shown in Fig. 7. Individual component calibration curves plotted in Figs. 8 and 9 are non-linear in the concentration range between 0.5 and 20 mg l^{-1}. An analyte interaction also exists for the solution mixtures because the fluorescence emission by rhodamine 6G, centered at 553 nm, falls under the rhodamine B absorption peak. The analyte interaction and the non-linearities noted in the individual component calibration plots give rise to a complex, non-linear two-component system. The rhodamine mixtures were split into fifty training patterns and thirty test patterns. One training set sample was discarded as an outlier based on an examination of the quadratic PCR training residual distribution [12]. The results obtained with the rhodamine mixtures were consistent with the performance advantage of neural networks that was demonstrated in the studies using simulated data. A comparison of SEP results in Table 6 reveals that the best results from a matrix regression method, PCR with cross-terms, contained 1.47 times the error of the 442sl neural network for rhodamine 6G, and 2.87 times the error for rhodamine B. A fairly high level of noise is present in the data since many absorbances above 2.2 were present in the mixture spectra. As the studies with synthetic data indicated, the presence of noise in the data produces a compression of the performance

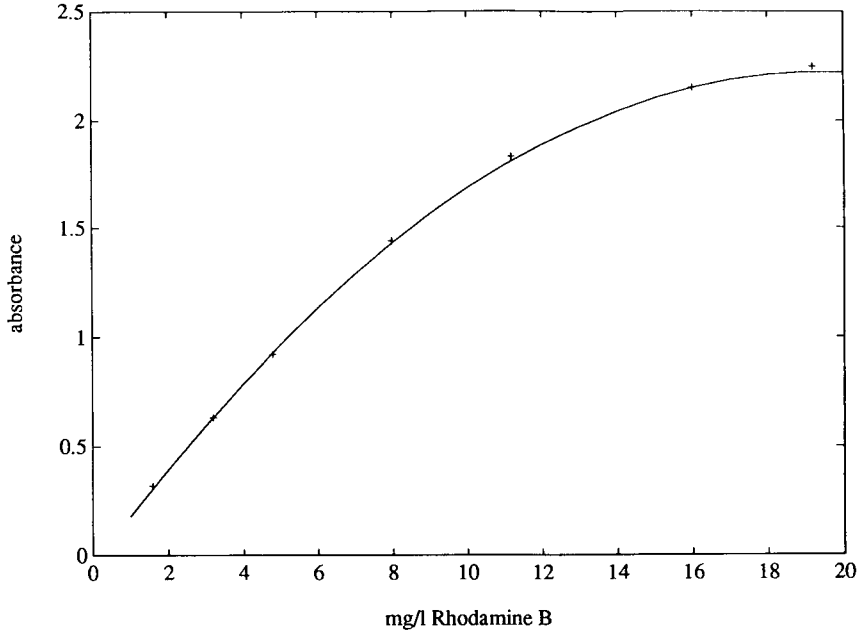


Fig. 9. Second-order polynomial fit of Rhodamine B calibration set. Absorbance data were taken at 550 nm.

TABLE 6

Calibration of rhodamine mixtures

Method	Error (%SEP)	
	Rhodamine B	Rhodamine 6G
Linear PLS	5.13	2.61
Polynomial PLS	4.89	4.27
Linear PCR	4.47	2.37
Quadratic PCR (+cross-terms)	4.14	2.19
1st order PCR (+cross-terms)	3.59	2.24
442sl NN	1.25	1.51

differences between methods. It is therefore probable that a lower noise level in the rhodamine data would increase the performance advantage of the neural network over the matrix regression.

Pattern classification

The asymmetric case in chemical pattern recognition consists of the separation of relatively tightly clustered class members from more diffusely distributed surrounding outliers. Both SIMCA and UNEQ may be used to separate outliers from class members in the asymmetric case by using statistical distribution information to develop class models. In contrast with SIMCA and UNEQ, feed-forward, layered neural networks separate classes by the construction of discriminant surfaces. The hidden nodes of a three-layer neural network function as hyperplane tests for the coarse separation of different classes of objects [22]. The output layer nodes are believed to refine class regions by using intersections of output layer hyperplane tests with those from the hidden layer. Pattern classification of test samples was determined by simply rounding the outputs of the neural network to the nearest integer in order to obtain binary-valued outputs. Each output node was used to designate a particular class with a rounded output of one indicating class membership, and a round output of zero indicating that the pattern was not a member of that class. For the asymmetric case the output from a single node was used to indicate a class member or an outlier.

The iris data set is a well-known classification problem with linearly separable classes [22]. Re-

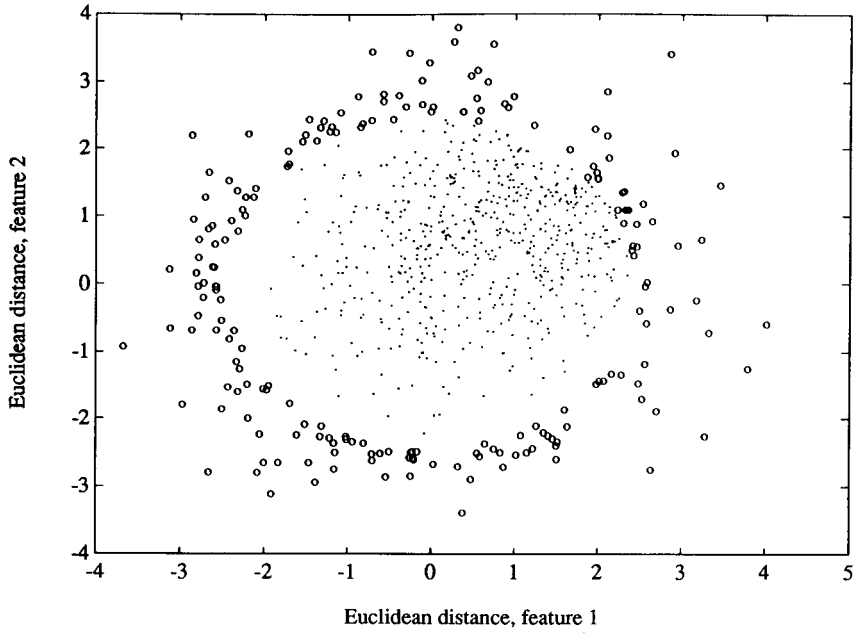


Fig. 10. Two-dimensional example of non-homogeneous asymmetric case. Class members denoted by (·), and outliers denoted by (○).

sults given in Table 7 shows that a 443ss network outperformed both UNEQ and SIMCA in the classification of the iris test set. Four objects were misclassified with SIMCA class modeling, and one was rejected as an outlier. With UNEQ class modeling, 2 objects were misclassified, and 8 were rejected as outliers. A 361ss neural network misclassified 3 objects, but rejected no objects as outliers.

TABLE 7

Classification of the iris data set

Method	Number correct per 75	Correctly classified (%)
443ss NN	72	96
UNEQ	65	86.7
SIMCA	70	93.3

Classification of class members in the asymmetric case amounts to fitting a spherical decision surface around a trivariate, normally distributed set of class members. Both homogeneous and non-homogeneous asymmetric case data were generated for examination. Training sets consisted of a full set composed of 150 class members and 100 outliers, and a smaller subset consisting of 75 class members and 50 outliers. An independent, universal test set was made up of 150 class members and 100 outliers. Only class members in each training set were used in the development of class models by the SIMCA and UNEQ class modelers, while neural network modeling required the use of both class members and outliers of the training set. An example of non-homogeneous, bivariate, normally distributed class members is shown in Fig. 10 along with homogeneous, bivariate, normally distributed outliers. The prediction errors caused by the non-homogeneity in the data are plotted in Fig. 11. The false class assignment of outliers, or false positives, occurs mainly in the first quadrant of the plot, and false rejection of class members, or false negatives, occur mainly in the second, third, and fourth quadrants of the plot. These error patterns arise from a shift of the UNEQ class decision surface caused by the shift of the training set centroid into the first quadrant.

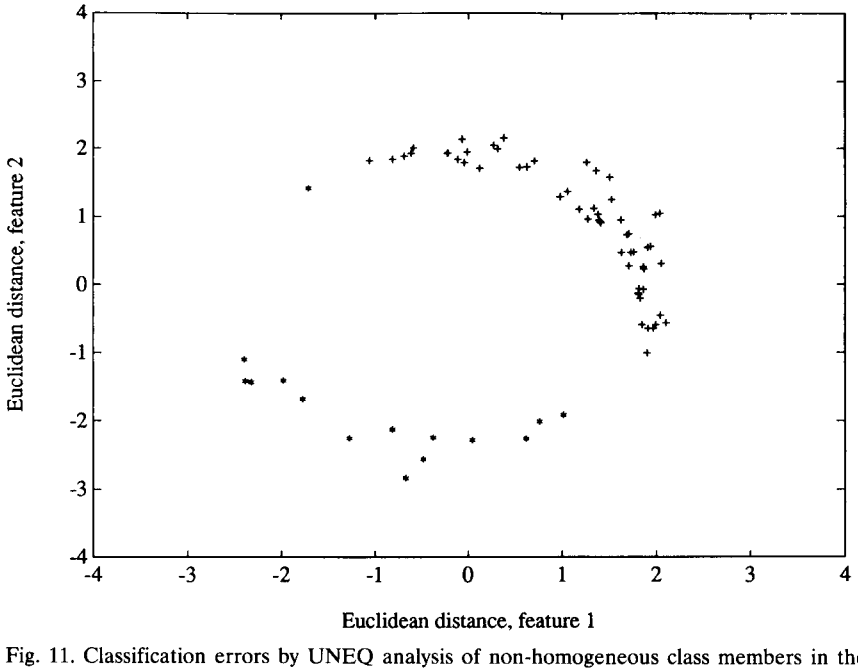


Fig. 11. Classification errors by UNEQ analysis of non-homogeneous class members in the asymmetric case: false positives (+), false negatives (*).

As expected, the UNEQ class modeler outperformed the SIMCA class modeler in classifying the homogeneous class-in-class data. The homogeneous, multinormal data structure and the 95% confidence decision boundary clearly favors the UNEQ method because UNEQ was developed under these assumptions. Results given in Table 8 support the conclusion that the 361ss neural network performs nearly as well as the UNEQ classifier, but UNEQ outperforms the neural network in the number of falsely rejected class members. This difference between UNEQ and the neural network is not surprising because in the training set the density of class members near the decision surface tends to be more diffuse, and neural network classification performance is known to depend on the density of training samples near the decision boundary [23]. The SIMCA classifier, when used with the usual SIMCA box with boundaries set by 95% confidence limits, was apparently not sensitive enough to detect outliers with the same multivariate normal distribution as class members. The performance of SIMCA would almost certainly improve relative to the others with an elliptical class boundary, but this modification was not investigated here, as the point of this research was an examination of the capabilities of neural networks.

Non-homogeneous class-in-class data were also examined with the three class modelers. This case is sometimes encountered in real data, and is expected to cause errors when the UNEQ classifier is applied to the data. The results obtained from the non-homogeneous data, presented in Table 9, are consistent with the expected increase in classification errors using the UNEQ classifier. The non-homogeneous nature of the class members in the training set appears to have no effect on the classification performance of the neural network, however, a result not unexpected because neural networks do not use distribution information to develop classification boundaries. As with the homogeneous data, all outliers were classified by SIMCA as class members.

From these studies, the performance of artificial, feed-forward, multi-layer neural networks has been shown to be comparable to traditional chemometric methods for two non-linear problems. While there is a decided time disadvantage to using artificial neural networks at present, the increasing pace of development of neural nets implemented in hardware (where their performance will no doubt improve significantly) suggests that they will be worth considering for dealing with linear and non-linear problems.

TABLE 8

Classification of the homogeneous asymmetric case

Method	Class members/ outliers [a]	False negatives	False positives
361ss NN	150/100	9	2
361ss NN	75/50	8	10
UNEQ	150/0	6	2
UNEQ	75/0	4	10
SIMCA	150/0	0	100
SIMCA	75/0	0	100

[a] In training set. Test set has 150 class members and 100 outliers in all cases.

TABLE 9

Classification of the non-homogeneous asymmetric case

Method	Class members/ outliers [a]	False negatives	False positives
361ss NN	150/100	5	5
361ss NN	75/50	6	10
UNEQ	150/0	4	25
UNEQ	150/0	5	21
SIMCA	150/0	0	100
SIMCA	75/0	0	100

[a] In training set. Test set has 150 class members and 100 outliers in all cases.

This research was supported by contract DAL03-91-C-0034 from the U.S. Army Chemical Research and Development Engineering Center.

APPENDIX

The modified delta rule for back-propagation using extended delta-bar delta learning is

$$\Delta w_{ji}(k) = -\eta_{ji}\frac{\partial E}{\partial w_{ji}} + \alpha_{ji}\Delta w_{ji}(k-1) \qquad \text{(A1)}$$

where α_{ji} = momentum coeff. for the connection weight between processing element i in layer i to processing element j in layer j; η_{ji} = learning coeff., weight ji; k = weight update index.

Adjustment of learning and momentum coefficients is based on the sign of the product function delta-bar delta

$$\bar{\delta}(k-1)\delta(k)$$

where $\bar{\delta}$, the weighted average gradient component, is defined as

$$\bar{\delta}(k-1) = (1-\Theta)[\delta(k-1) - \Theta\delta(k-2)] \tag{A2}$$

and where $\delta(k-1)$ = gradient component, $(\partial E/\partial w_{ji})$, for the weight ji at weight correction index $k-1$; Θ = convex weighting factor (usually 0.3 to 0.7).

The learning coefficients are updated by

$$\begin{aligned}
&\Delta\eta(k) = \kappa_\eta \exp\left(-\gamma_\eta |\bar{\delta}(k)|\right) \\
&\quad \text{if } \bar{\delta}(k-1)\delta(k) > 0 \\
&\Delta\eta(k) = -\phi_\eta \eta(k) \quad \text{if } \bar{\delta}(k-1)\delta(k) < 0 \\
&\Delta\eta(k) = 0 \quad \text{otherwise}
\end{aligned} \tag{A3}$$

where κ_η = learning rate increase scale factor; ϕ_η = learning rate decrease scale factor; γ_η = exponential factor for learning rate increase.

The momentum coefficients are updated by

$$\begin{aligned}
&\Delta\alpha(k) = \kappa_\alpha \exp\left(-\gamma_\alpha |\bar{\delta}(k)|\right) \\
&\quad \text{if } \bar{\delta}(k-1)\delta(k) > 0 \\
&\Delta\alpha(k) = -\phi_\alpha(k) \quad \text{if } \delta(k-1)\delta(k) < 0 \\
&\Delta\alpha(k) = 0 \quad \text{otherwise}
\end{aligned} \tag{A4}$$

where κ_α = momentum coeff. increase scale factor; ϕ_α = momentum coeff. decrease scale factor; γ_α = exponential factor for momentum coeff. increase.

REFERENCES

1 J.R. Long, V.G. Gregoriou and P.J. Gemperline, Anal. Chem., 62 (1990) 1791.
2 C. Borggaard and H.H. Thodberg, Anal. Chem., 64 (1992) 545.
3 P.J. Gemperline, J.R. Long and V.G. Gregoriou, Anal. Chem., 63 (1991) 2313.
4 M. Bos and H.T. Weber, Anal. Chim. Acta, 247 (1991) 97.
5 J.R.M. Smits, L.W. Breedveld, M.W.J. Derksen, G. Kateman, J. Snoek and J.W. Hofstraat, Anal. Chim. Acta, 258 (1992) 11.
6 R.J. Fessenden and L. Gyorgyi, J. Chem. Soc., Perkin Trans. 2, (1991) 1755.
7 Y. Ito, Neural Networks, 5 (1992) 105.
8 N.B. Vogt, Chemom. Intell. Lab. Syst., 7 (1989) 119.
9 S. Wold, N. Kettaneh-Wold and B. Skagerberg, Chemom. Intell. Lab. Syst., 7 (1989) 53.
10 S. Wold and M. Sjostrom, in B.R. Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society, Washington, DC, 1977, 243.
11 M.P. Derde and D.L. Massart, Anal. Chim. Acta, 184 (1986) 33.
12 T.B. Blank and S.D. Brown, Anal. Chem., submitted for publication.
13 T.B. Blank and S.D. Brown, in preparation.
14 D.E. Rummelhart, G.E. Hinton and R.J. Williams, in D.E. Rummelhart and J.L. McClelland (Eds.), Learning Internal Representations by Error Propagation, Distributed Processing, Vol. 1, MIT Press, Cambridge, MA, 1986 p. 318.
15 P.A. Jansson, Anal. Chem., 63 (1991) 357A.
16 J. Zupan and J. Gasteiger, Anal. Chim. Acta, 248 (1991) 1.
17 R.L. Watrous, Learning Algorithms for Connectionist Networks: Applied Gradient Methods of Nonlinear Optimization, Proc. of the 1st ICNN, 2 (1987) 619.
18 A.A. Minai and R.D. Williams, Acceleration of Back-Propagation through Learning Rate and Momentum Adaptation, International Joint Conference on Neural Networks, 3 (1990) 676.
19 T. Tollenaere, Neural Networks, 3 (1990) 561.
20 B. Batchelor, Practical Approach to Pattern Recognition, Plenum Press, New York, 1974.
21 E.B. Baum and D. Haussler, Advances in Neural Information Processing Systems 1, Morgan Kaufmann, Palo Alto, CA, 1989, p. 81.
22 R.P. Brent, IEEE Trans. Neural Nets., TNN2 (1991) 346.
23 K.G. Mehrotra, C.K. Mohan and S. Ranka, IEEE Trans. Neural Nets., TNN2 (1991) 548.

Analytica Chimica Acta, 277 (1993) 289–295
Elsevier Science Publishers B.V., Amsterdam

Artificial neural networks as a multivariate calibration tool: modeling the Fe–Cr–Ni system in x-ray fluorescence spectroscopy

A. Bos, M. Bos and W.E. van der Linden

Department of Chemical Analysis, University of Technology Twente, P.O. Box 217, 7500 AE Enschede (Netherlands)

(Received 3rd September 1992)

Abstract

The performance of artificial neural networks (ANNs) for modeling the Cr–Ni–Fe system in quantitative x-ray fluorescence spectroscopy was compared with the classical Rasberry–Heinrich model and a previously published method applying the linear learning machine in combination with singular value decomposition. Apart from determining if ANNs were capable of modeling the desired non-linear relationships, also the effects of using non-ideal and noisy data were studied. For this goal, more than a hundred steel samples with large variations in composition were measured at their primary and secondary K_α and K_β lines. The optimal calibration parameters for the Rasberry–Heinrich model were found from this dataset by use of a genetic algorithm. ANNs were found to be robust and to perform generally better than the other two methods in calibrating over large ranges.

Keywords: Multivariate calibration; Principal component analysis; X-ray fluorescence spectrometry; Artificial neural networks; Chromium; Iron; Neural networks; Nickel; Steel

Artificial neural networks (ANNs) have gained much focus in recent years due to their ability to 'learn' arbitrary non-linear relationships between input and output spaces. Applications of ANNs in current literature can be divided in two distinct classes: as pattern classification systems, and as 'function approximation' systems. The latter is commonly referred to in the field of chemometrics as multivariate calibration.

The application of instrumental techniques in analytical chemistry usually implies the need for calibration in order to obtain accurate results. When the relationship (or model) between measured signals and concentrations is known to be linear, or can be made linear through transformation, standard methods can be applied for calibration of which principal components regression analysis (PCR) and singular value decomposition (SVD) are best known. In other cases a non-linear model must first be determined of which the characteristic parameters can be found through appropriate calibration methods. The simultaneous quantitative determination over large ranges of iron, nickel and chromium concentrations in stainless-steel samples with x-ray fluorescence spectroscopy is such a case. Due to strong interelement effects between these components the concentrations are not proportional to their relative signals and a non-linear model is required.

In this paper, three empirical approaches to solve this problem are compared: the approach proposed in the paper by Rasberry and Heinrich

Correspondence to: W.E. van der Linden, Department of Chemical Analysis, University of Technology Twente, P.O. Box 217, 7500 AE Enschede (Netherlands).

[1] which is based on empirically modeled interelement effects, a linear-based approach using singular value decomposition on the set of signals and their cross-products, and an approach in which the measured signals are used directly as inputs to a neural network which provides the desired concentrations at its outputs.

THEORY

Measured intensities in x-ray fluorescence analysis are usually not proportional to the concentrations in the specimen. Apart from inhomogenous and particulate samples, this can occur when the x-ray emission of the analyte is significantly affected by concentration variations of the other elements in the sample. These interelement effects depend on the mass fractions of the elements in the sample, and the relationships of their absorption coefficients for the primary and secondary radiation. Also secondary fluorescence can occur. These effects are illustrated in Fig. 1. In order to obtain accurate results, a non-linear calibration model is required. These models can be divided in two categories, one based on equations derived from theoretical principles and physical constants, and the other based on empirical calibration using standards. In the Cr–Ni–Fe system found in most alloys, strong interelement effects occur. The FeK_α is absorbed by chromium, and secondary fluorescence is induced in iron by the NiK_α. The NiK_α is absorbed in turn by chromium and iron, whereas iron and nickel both induce secondary fluorescence in chromium. In the following sections, three approaches for modeling this complex system are explained and compared.

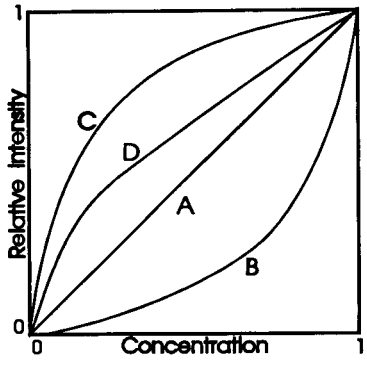


Fig. 1. Types of interelement effects in x-ray fluorescence spectroscopy: (A) linear calibration curve; (B) preferential absorption by matrix; (C) preferential absorption by analyte; (D) secondary fluorescence.

The Rasberry–Heinrich model

In this model the following empirically derived relationship is proposed

$$\frac{C_i}{R_i} = 1 + \sum_{k \neq i} A_{ik} \cdot C_k + \sum_{k \neq i} \frac{B_{ik} \cdot C_k}{1 + C_i} \qquad (1)$$

in which C_i and C_k denote the concentrations (fractions) of components i and k, R_i the relative intensity of component i, i.e., the measured intensity divided by the intensity found for the pure element, and A_{ik} and B_{ik} are calibration coefficients compensating for absorption and secondary fluorescence effects respectively. In their paper, the authors apply only the coefficients of the *dominant* effect, thus leading to *six* coefficients in a ternary system instead of the twelve possible. The calibration coefficients have to be determined empirically from a set of samples with known compositions. After calibration, unknown concentrations can be calculated by setting the initial concentrations equal to the relative intensities found and then iteratively recalculating new concentrations through Eqn. 1 until convergence is achieved. A flow chart of the algorithm can be found in [1]. The calibration can either be done by solving the set of equations or by graphical methods [1], but this would require too much effort when using a leave one out approach. Therefore, in this work a genetic algorithm (GA) is applied as combinatorial optimization technique for finding the required calibration constants. A full description of GAs can be found in [2,3], and an application to analytical chemistry in [4]. Here only a short description of the characteristics is given.

A GA is a stochastic combinatorial optimization method based on an artificial simulation of the Darwinistic evolutionary process which pro-

vides a robust and efficient method for searching very large parameter spaces. Randomly initialized sets of parameters are mapped onto 'chromosomes', and an evolutionary process is started in which all individuals (sets of parameters) compete for reproduction through a 'fitness function'. By using crossover functions, new generations are capable of combining beneficial subparts of the chromosomes to create better fitting individuals, thus 'evolving' to an optimal solution. Mutation is applied to keep diversity in the gene pool. Whereas the choice of a correct cooling curve is essential in the case of the comparable 'simulated annealing' method, the choice of pool size, fitness function and crossover and mutation probabilities is essential for GA's.

A semi-linear approach

As mentioned before, the concentrations are not directly proportional to the measured intensities, thus prohibiting the application of linear calibration. This can be overcome by expanding the data set with quadratic and cross product terms of the measured intensities as has been demonstrated earlier [4]. After measuring a set of samples, a relationship can be made for each element of interest as in Eqn. (2).

$$\mathbf{A} \cdot \vec{x} = \vec{c} \tag{2}$$

The matrix **A** contains in each row the measured intensities and their quadratic and cross product terms, and the vector $\vec{c}$ in the same row the corresponding concentration. An optimal solution vector $\vec{x}$ can then be determined through means of singular value decomposition [5].

Neural networks

Artificial neural network theory supplies us with a general adaptive model for *learning* an arbitrary non-linear mapping from an input space to an output space. This is done by simulating a network topology and then presenting it with a database of *examples* and applying a learning rule. Through the learning rule, the network will adapt and learn from the examples to respond correctly to its environment. Especially the perceptron-based backward error propagation networks, or backpropagation networks, are widely

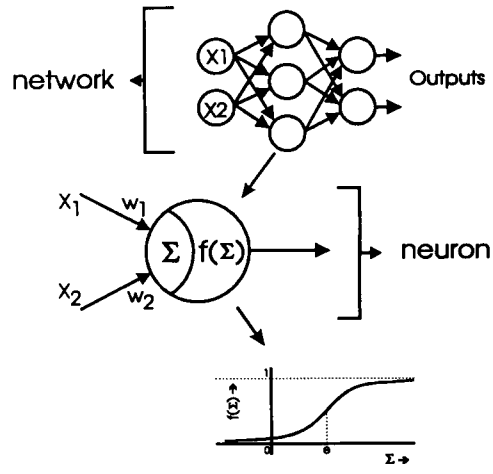


Fig. 2. A typical backpropagation network consisting of sigmoid neurons.

used in many areas and as such are the focus of intensive research and numerous papers. A short recapitulation of this type of network and its properties will be given here. Descriptions of backpropagation networks applied to analytical chemistry have been given in earlier papers [6,7]. For a more explicit description refer to Ref. 8.

Backpropagation networks are mainly characterized by the type of units (neurons) used, by the topology [interconnections and their weights (or strengths)] of the network, and by the learning rule applied. The basic units, of which the 'sigmoid neuron' is best known, perform a weighted processing of the input signals, which are either pattern inputs or outputs of other neurons. This is followed by a non-linear scaling or squashing function as is illustrated in Fig. 2 and Eqns. 3a–3b.

$$net = \sum_i x_i \cdot w_i \tag{3a}$$

$$f(net) = \frac{1}{1 + e^{-net + \theta}} \tag{3b}$$

The biases (θ_i) for each neuron are usually replaced by extra weights to an imaginary constant input of magnitude 1 for easier programming. A network is built by combining these units into layers, and then applying at least one layer, a

so-called 'hidden' layer, of neurons between the input and the output layer of the network. The non-linearity of the squashing function and the occurrence of at least one hidden layer are required to allow a network to learn a non-linear mapping. The free variables in a network are the weights, or interconnection strengths, usually represented as floating point numbers. The goal of the learning rule is to find a set of weights for a given network topology by initializing the weights to random values and then repeatedly adapting them until the network produces the correct outputs for each set of inputs in the examples. In general there are many, if not unlimited, possible sets of weights leading to the same global behavior of the network. Which one is found depends on the random initialization of the weights before the training starts and the implementation of the training phase.

The iterative process of adapting the weights is usually a time-consuming task. The adaptation of the weights in the original backpropagation algorithm is done by first order gradient descent, with a fixed step size parameter chosen by the experimenter, which converges very slowly near the optimum due to the small gradient. This effect can be countered by introducing an ad hoc momentum factor. To obtain accurate *quantitative* results, one must continue training for very long times. In recent years, several adaptations of the conjugate gradient methods, which apply second order information in determining step directions and sizes, have been proposed for backpropagation networks [9,10]. Of these, the 'scaled conjugate gradient' (SCG) method as proposed by Möller [10] has been implemented by us with good results. Möller's proposal applies a Levenberg–Marquardt approach in scaling the step size as opposed to the more common, and computationally intensive, line search found in other papers. The resulting algorithm, although rather complex, requires no external or empirical parameters to be set like the learning rate and momentum in the original backpropagation. The high rate of convergence, theoretically quadratic near the optimum for conjugate gradient methods, enables achieving accurate quantitative results in acceptable calculation times.

EXPERIMENTAL

The data set

About 130 stainless-steel samples of which the composition varied widely were provided to us by the faculty of Mechanical Engineering at our university. Apart from iron, nickel and chromium, also relatively large amounts of manganese, cobalt and vanadium were present, as well as traces of many other elements. Of the samples, a varying number of components had been determined in the past through wet-chemical analysis with unknown accuracy. All samples had polished surfaces and were measured quantitatively in a Philips PW1480 spectrometer under control of the Philips X44 software running on a PDP-11 minicomputer. The x-ray tube had a chromium anode, therefore an aluminum filter of 0.085 mm thickness was used when measuring near the chromium fluorescence lines to compensate for the chromium radiation from the tube itself. For all elements an LiF-200 crystal and fine collimator were used. Nickel and the primary K_α of chromium were detected by flow counter and scintillator, and the others by flow counter alone. To measure the fluorescence of the elements with low concentrations sufficiently, the tube was operated at 50 kV and a current of 50 mA. All primary and secondary K_α and K_β lines of all major components were measured, although only parts of these data were used during the comparison tests. Also intensities of pure samples of all elements found in the steel samples were determined. When possible directly from the pure element itself, in other cases through compound standards and calculations. The high operating current of the tube saturated the detector when determining the strong fluorescence of the primary lines of pure iron and nickel, therefore for these elements a linear calibration at lower currents was done and an extrapolation to 50 mA made.

Of the acquired data, those samples with obvious errors such as total compositions of much more than 100% and those of which the rest (the components which are not iron, nickel or chromium) exceeded 5% of the total were taken out. The remaining 101 samples contained com-

TABLE 1

Concentrations found in the steel samples

	Min. (%)	Max. (%)
Iron	50.48	98.17
Nickel	0.02	21.80
Chromium	0.04	29.20
Rest	0.72	4.03

positions varying over a large range, as illustrated in Table 1.

Procedures and parameters

All models were tested using a 'leave one out' approach to gain maximum information from the available data. In this approach one pattern is removed from the data set and the model is calibrated with the remaining patterns. The pattern taken out is then evaluated as an 'unknown' and the errors are calculated. Then the pattern is put back and the whole process is started anew for the next one in the data set. In this manner each pattern serves as an unknown as well as a calibration pattern.

For the Rasberry–Heinrich model, the Cr$K_{\alpha 1}$, the Ni$K_{\alpha 1}$, the Fe$K_{\alpha 1}$ and respective concentrations were selected from the original data set. Two leave one out experiments were done, one as a Cr–Ni–Fe ternary system, and in the other the remaining components were treated as an imaginary fourth component 'X' as suggested in the original paper [1]. In the former case, the six coefficients as suggested by the authors, and in the latter case six extra coefficients for absorption and secondary fluorescence effects of X were determined. This was done because no dominant effect of X was known. In the algorithm, the fraction of X was continuously calculated as one minus the fractions of Cr, Ni and Fe. If the fraction of X became negative, it was set to zero. As a check for the application of the genetic algorithm in finding the calibration constants, a test run was performed with the data as published in the original article [1].

To decrease the calculations to manageable proportions for the genetic algorithm, first an approximation of the calibration coefficients was derived from the whole data set using a population of 60 individuals with 16 gray-coded bits allocated for each coefficient scaled in the interval [−3.0, 3.0]. The algorithm was allowed to continue until no more improvement was observed. Some random noise was added to the coefficients found and they were then used as a base in the leave one out method. In this method, a population size of 40 was used, and offsets from the previously found base-coefficients were gray-coded as 8 bits scaled in the interval [−0.5, 0.5], resulting in a resolution of approximately 0.004 for each coefficient. The crossover rate was set to a chance of 0.2, and the mutation rate to a probability of 0.01 per bit location. After some testing, the number of generations was fixed at 50, allowing full convergence for all samples in the data set.

For the semi-linear test, the same signals as for the Rasberry–Heinrich model were used expanded with all possible quadratic and cross product terms. Singular value decomposition was used with a varying number of terms in the diagonal matrix to find optimal solution vectors $\vec{x}$ for each given set of training examples. Also the inclusion of intensities for the trace elements (Mn, Co and V) were tried, but this led to slightly worse results than with the intensities for the three main elements alone.

All neural nets were of the fully connected feed forward type employing the sigmoid function as squashing function. Primary K_α and K_β of Cr, Ni, Fe, Mn, Co and V were used as input signals. All inputs were scaled in the interval [0.1, 1.0] and the outputs were scaled in the interval [0.1, 0.9]. Scaling of the outputs is required as the sigmoid function is asymptotic near 0 and 1. The scaling of the inputs was not required in this case, especially as the signals were relative, but is an automatic feature of our software. Preliminary experiments, however, showed no increase in performance when using unscaled signals for training. All networks were trained for 15 000 epochs using the scaled conjugate gradient algorithm allowing maximum convergence. Different networks were used for each of the three components to simplify optimization and a range of network topologies were tested. The best results were used for the comparison.

TABLE 2

Comparison of calibration constants found by the genetic algorithm (GA) for the dataset and those as published by Rasberry and Heinrich [1]
(A denotes absorption and B denotes secondary fluorescence effects)

Effect	Type	Coefficients	
		[1]	GA
Cr → FeK_α	A	2.10	2.11
Cr → NiK_α	A	1.20	1.21
Fe → NiK_α	A	1.71	1.75
Fe → CrK_α	B	−0.46	−0.40
Ni → FeK_α	B	−0.47	−0.44
Ni → CrK_α	B	−0.27	−0.23

RESULTS AND DISCUSSION

In Table 2 the calibration constants found by the genetic algorithm for the dataset of binary and ternary samples as originally published by Rasberry and Heinrich [1] are given, as well as the constants they found. Several runs with different random starting populations converged invariably to the same set of coefficients. The mean error for the set of five unknowns are compared in Table 3. Although with the GA slightly different constants are found, the results in predicting the unknown samples are better.

With the dataset of steel samples, however, a distribution of coefficients was found depending on which sample was left out at the time. The distributions found are given in Tables 4 and 5 for the ternary and quaternary systems respectively. These values indicate the sensitivity of the model for data containing relatively large noise amplitudes. The standard deviations for X in the quaternary system are relatively high due to the small and strongly varying concentrations of the constituent trace elements. Of these, the effects of X on the NiK_α and of nickel on the CrK_α appear to have no statistical significance.

TABLE 3

Results for the set of unknowns compared

	Mean error	
	[1]	GA
Chromium	0.0027	0.0016
Nickel	0.0030	0.0025
Iron	0.0045	0.0034

TABLE 4

Distribution of coefficients found for the actual data set (A denotes absorption and B denotes secondary fluorescence effects)

Effect	Type	Coefficient	
		Mean value	S.D.
Cr → FeK_α	A	2.60	0.004
Cr → NiK_α	A	1.31	0.240
Fe → NiK_α	A	2.03	0.011
Fe → CrK_α	B	−0.50	0.094
Ni → FeK_α	B	−0.96	0.078
Ni → CrK_α	B	−0.34	0.025

In the case of the neural network approach, a number of network topologies were tried for all elements. A topology with four neurons in the hidden layer proved to be sufficient for all three elements, and was used in the leave one out method. However, no decrease in performance due to 'overfitting' of the data was detected when larger topologies were used.

The best results for each method of all leave one out experiments are listed for each primary element separately in Table 6. The mean values of the absolute (fractions) and relative (percentages) errors are given. The latter gives rise to very high errors at the low concentrations of chromium and nickel in some of the samples and

TABLE 5

Distribution of coefficients found for the quaternary system (A denotes absorption and B denotes secondary fluorescence effects)

Effect	Type	Coefficient	
		Mean value	S.D.
Cr → FeK_α	A	2.53	0.076
Cr → NiK_α	A	1.42	0.260
Fe → NiK_α	A	1.95	0.095
Fe → CrK_α	B	−0.46	0.020
Ni → FeK_α	B	−1.39	0.181
Ni → CrK_α	B	−0.14	0.125
X → CrK_α	A	1.38	0.216
X → CrK_α	B	1.02	0.248
X → NiK_α	A	0.08	0.281
X → NiK_α	B	0.86	0.277
X → FeK_α	A	0.70	0.131
X → FeK_α	B	0.56	0.256

TABLE 6

Errors of the different methods for chromium, nickel and iron [NN denotes the neural network, SLM the semi-linear method R&H(3) and R&H(4) the ternary and quaternary Rasberry–Heinrich models, and linear the pure linear calibration. Absolute errors are expressed in fractions, relative errors in percentages]

Method	Chromium		Nickel		Iron	
	Abs.	Rel. (%)	Abs.	Rel. (%)	Abs.	Rel.
NN	0.0023	20.63	0.0015	29.03	0.0053	0.68
SLM	0.0022	11.21	0.0020	49.46	0.0064	0.82
R&H (3)	0.0062	11.26	0.0028	32.14	0.0453	5.46
R&H (4)	0.0034	7.89	0.0020	30.98	0.0245	2.99
Linear	0.0087	136.03	0.0039	128.32	0.0363	4.49

serves to illustrate the performance of the various models for these low concentrations. As an illustration of the non-linearity of the measured system, the results of a purely linear calibration are also listed in the tables.

In the case of chromium, the neural net performs slightly worse than the semi-linear model, but much better than the Rasberry–Heinrich model. The latter, however, produces better results for low concentrations as can been seen from the relative errors. In the cases of nickel and iron, the neural net outperforms all other models. Surprising are the relatively bad results of the Rasberry–Heinrich models for iron. This in our opinion can be ascribed to the large levels of noise present in the data set. For all three elements, the application of a quaternary system for the Rasberry–Heinrich model has shown to improve the results a great deal.

Conclusion

In the presented case of determining the composition of steel over large ranges, backpropagation neural networks have shown to be capable of robust multivariate calibration. They can supply accurate quantitative results and have outperformed two other previously known calibration models for this problem. Also, no tendency was detected to overfit the data set when very large topologies were tested or extreme long training times were applied.

The use of the SCG algorithm has led to faster converging networks and has decreased the total amount of calculation time required considerably. Also fast convergence was observed when two hidden layers were used during the search for an optimal topology. In these cases the standard backpropagation algorithm tends to be very slow.

REFERENCES

1 S.D. Rasberry and K.F.J. Heinrich, Anal. Chem., 46 (1974) 81.
2 D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA, 1989.
3 J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, 1975.
4 M. Bos, Anal. Chim. Acta, 166 (1984) 261.
5 W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, New York, 1989.
6 M. Bos, A. Bos and W.E. van der Linden, Anal. Chim. Acta, 233 (1990) 31.
7 A. Bos, M. Bos and W.E. van der Linden, Anal. Chim. Acta, 256 (1992) 133.
8 D.E. Rumelhart and I.L. McClelland, Parallel Distributed Processing, Vols. 1 and 2, MIT Press, Bradford, 1986.
9 E.M. Johansson, F.U. Dowla and D.M. Goodman, Lawrence Livermore National Laboratory, 1990, Preprint UCRL-JC-104850.
10 M.F. Möller, Neural Networks, in press.

Analytica Chimica Acta, 277 (1993) 297–304
Elsevier Science Publishers B.V., Amsterdam

Hypermedia tools for the interpretation of mass spectra

T. Brodmeier, A. Gloor, M. Cadisch, R. Bürgin and E. Pretsch
Department of Organic Chemistry, Swiss Federal Institute of Technology, CH-8092 Zurich (Switzerland)

(Received 3rd September 1992)

Abstract

A Hypermedia application has been developed for the interpretation of molecular spectra. It supports the work of the chemist by providing reference data, reference spectra and computational tools. The pieces of information that are available for the interpretation of mass spectra are described.

Keywords: Mass spectrometry; Hypermedia application; Interpretation of mass spectra

Owing to the difficulties connected with the prediction of spectra from a chemical structure, the interpretation of molecular spectra, including mass spectra, is mainly an empirical procedure. It is usually based on various kinds of reference data, reference spectra, mass balances and other calculations. A huge amount of proven fragmentation mechanisms for a wide range of compound types has been collected in recent decades. Although they cannot be applied directly to other structures, their knowledge may be of great help when spectra of analogous compounds are interpreted. Collections of such mechanisms are only available in printed media [1,2]. Generalizations are possible for selected compound classes, so that heuristic rules of typical fragmentation processes can be derived, which may be of importance in specific cases. A further source of knowledge is the inspection of reference spectra in printed media [3] or preferably as databases [4]. Such databases together with various search programs and further tools for interpreting the results [5] are available as part of the software of instrument computers or as stand-alone spectroscopic data banks [6–8]. The databases contain up to over 100 000 spectra of variable quality [9]. No database of interpreted spectra or of fragmentation mechanisms is available.

Besides inspecting such reference data, the chemist often has to perform trivial mass balance or book-keeping calculations to obtain the possible elemental compositions of the fragments of a given mass. Such calculations including those of isotope signal abundances for a selected elemental composition have in common that no heuristic knowledge needs be used for them.

Recently SpecTool, a Hypermedia application for spectrum interpretation, has been described [10–12]. Hypermedia is an information management technology that links text, graphics, sound and computer programs in an associative manner [13–15]. It allows the user to navigate through the pieces of information in a non-linear way. SpecTool, the first Hypermedia application in analytical chemistry, is based on HyperCard, the most widely available Hypermedia environment for Macintosh computers. Reference data, selected prototype spectra and programs are collected in the nodes of SpecTool. The nodes are connected with a complex network of links. Typically 50

Correspondence to: E. Pretsch, Department of Organic Chemistry, Swiss Federal Institute of Technology, CH-8092 Zurich (Switzerland).

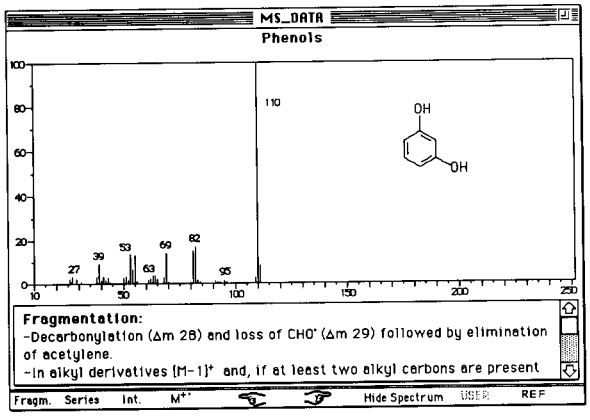


Fig. 1. MS-data card with display of spectrum.

nodes can be directly accessed in one step from any position in the system. This complex network can be dynamically organized in various logical structures allowing one to browse through data, spectra and programs. Further, it is possible to switch to related pieces of information in various other spectroscopic techniques.

DATA AND SPECTRA

The data include numerical constants such as isotope abundances, which are collected in a table together with various presentation tools allowing a fast look-up and both numerical and graphical presentation [11]. Further, heuristic rules of typical fragmentation pathways of monofunctional compounds are collected in over 100 nodes. In most instances, prototype spectra can be inspected together with the rules (Fig. 1). As a further class of data, proven fragmentation pathways of selected compounds are also included. These parts of the system have been described elsewhere [11,12].

COMPUTATIONAL TOOLS

All tools provided with the system are for straightforward calculations without using any heuristics.

Molform

A program described earlier [16] allowing the calculation of all possible molecular formulae for a user-defined mass range has been interfaced to the system. An especially fast algorithm is used with high-resolution spectra. The user may enter further constraints regarding element ranges (obtained, e.g., from NMR spectra), weight percentages of elements (obtained from elemental analysis data) or double bond equivalents.

Isotope pattern

A recently published fast algorithm [17] has been implemented as a Pascal program and interfaced to the system. On entering the elemental composition, the calculated isotope abundances are displayed both numerically and graphically (Fig. 2).

Peak-Explorer

A further tool provides support during the interpretation of mass spectra. Its main ability is the calculation of all possible elemental compositions for a single mass signal. Because this kind of question often results in a numberless amount of solutions, the possibility of constraints has been introduced. Further, there are options for sorting the results according to various criteria in addition to saving and loading them. Not only the calculated elemental compositions but also some available literature information [18] belonging to the chosen mass are displayed. These items serve as starting points for branching into data collec-

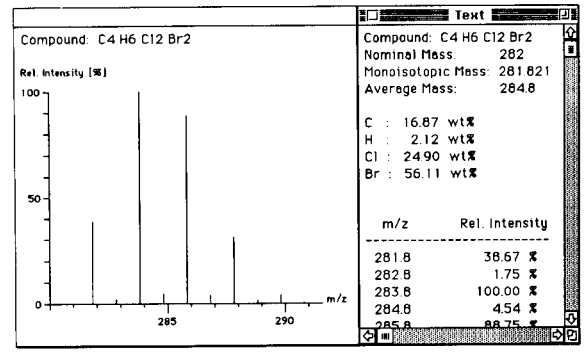


Fig. 2. Isotope pattern output.

tions of SpecTool. Thus reference spectra for comparison purpose are accessible.

In contrast to most other parts of SpecTool, this instrument combines the properties of applications and data objects. This hybrid constitution could ideally be taken into account by using a suitable programming language type for each aspect. Thus the user interface and all data are located in a HyperCard stack, whereas some more demanding arithmetic processes take place in compiled Pascal code, which can be called from the HyperCard environment as an external procedure.

As can be seen in Fig. 3, the user can specify constraints concerning double bond equivalents (DBE). This is possible by defining the allowed range or by excluding hypervalent species. Hypervalent species are those with DBE < -0.5, i.e., species protonated more than once. Thus NH_5 and CH_6 will not succeed, whereas NH_4 and CH_5 are interpreted as valid elemental compositions.

The central input parameters are the signal mass and its maximum deviation. The signal mass can be entered as nominal or exact mass. The user has to consider that all calculations are executed with exact isotopic masses. Therefore, nominal masses should be used with care as input values, especially for heavier species, otherwise the increasing difference between nominal and monoisotopic masses will inevitably lead to wrong results.

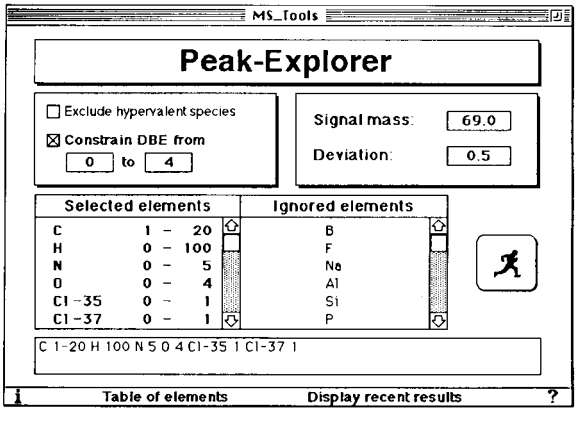


Fig. 3. Main card of Peak-Explorer.

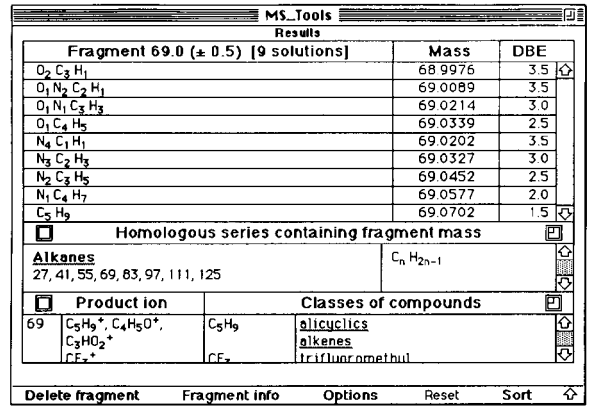


Fig. 4. Result card of Peak-Explorer.

The allowed ranges for the numbers of elements can be entered with an input string. The list of all possible elements is shown in a field to remind the user that different isotopes (e.g., Cl-35 and Cl-37) and different valences [e.g., S(2) and S(6)] of an element are distinct input items. The table of elements can be edited by the user. The definition of a group of elements as super atom is thus also possible. A dialogue box will alert the user in case of syntax errors.

The upper part of the display of results (Fig. 4) represents the calculated elemental compositions with their monoisotopic masses and double bond equivalents (DBE). This main list is followed by additional literature information [18] if available. First, an extract of the mass correlation table is quoted. It contains four columns: signal mass, frequent elemental compositions for product ions and for neutral losses corresponding to the specified mass and classes of compounds typically exhibiting signals at the specified mass or mass difference. In a second field, frequent homologous ion series matching with the chosen mass are shown. Both of these data collections may be used for navigation purposes to the corresponding entries of SpecTool's MS data stack.

Further options (buttons or pop-up menus on the bottom part of the card) are available for deleting fragments, changing displays, saving and loading results and sorting.

Sorting is possible based on elements, DBE and fragment masses. In addition to these sort procedures it is possible to display only parts of the list: either "Only non-radicals" (i.e., only elemental compositions with non-integral DBE), or "Only radicals" (i.e., only elemental compositions with integral DBE).

As on all levels of SpecTool, clicking at the title field makes a navigation pop-up menu appear. This allows the user to move upwards along the tree of the hierarchical table of contents or simply to obtain some orientation information when lost. Macintosh's System 7 feature "Balloon Help" is fully supported by this tool.

Implementation. One of the most powerful features of HyperCard is the extension of its built-in commands and features by adding compiled code as external resources. These objects are pasted into the resource fork of a HyperCard stack using a resource editor, whereupon they can be called by a specified name from any HyperCard script. Without this kind of reinforcement, it would be impossible for HyperCard to execute large-scale procedures within reasonable time.

The Peak-Explorer tool uses four code resources of the described type: one for the evaluation of input strings on the main card, one for the calculation of elemental compositions and two for sorting. Apart from that, the application utilizes two font resources and one sound resource.

The main task of this tool is to solve a combinatorial problem as follows: find for a given set of elements systematically all elemental compositions that lie within a defined range. In a continuous trial-and-error process, thousands of elemental compositions are checked for conformity with the required mass interval. The algorithm always starts with the heaviest element, estimates its maximum amount and tries to fill the remaining gap with more light-weight species. When the end of the set is reached, the last but one is decremented step by step, always checking for successful compositions, and so on. Thus the algorithm returns to the heaviest element remaining in contest, decrements it by one and the whole exercise starts again. As soon as that heaviest species is set to zero, all its possible combinations have been examined and it no longer comes into question during further cycles. Recursive programming had to be avoided because of memory size restrictions for external routines.

TIMS

TIMS is another tool supporting the interpretation and rationalization of mass spectra. In essence it performs straightforward calculations of the masses obtained by simple fragmentations of a molecule. Thus the spectroscopist can concentrate on those parts of the mass spectrum that cannot be explained by simple fragmentations.

Such a book-keeping program was first reported by Kubinyi [19]. MSBreak is a program that can propose structural elements interpreting fragmentations and intensities of isotope signals and also using some heuristic rules. InterpMS [20], an earlier version of TIMS, was then developed. Another similar program is MASSPEC [21], running on personal computers and VAX workstations. It is also a graphics-based program that correlates a mass spectrum with a structure. It can be used to determine sequences in short peptides and supports interpretation of fast atom bombardment and tandem mass spectra.

As it does not contain any heuristic rule, the information one can obtain from TIMS is absolutely reliable. It works with low-resolution mass spectra, but supports high-resolution information for calculated elemental compositions. One strong point of TIMS is its graphic user interface; in fact, all information available in text windows can be displayed or elaborated in graphics windows containing the spectra and the structure proposed by the user.

TIMS can be used with greatest benefit if one can provide both a proposed structural formula and a mass spectrum. The structure must be entered in connection table format, which can be easily generated by most of the common chemical structure drawing programs. The user should avoid introducing explicitly bonded protons, because TIMS considers the single atoms as super atoms carrying their protons and fills up free valences with protons. However, the formula may contain a user-specified super atom (defined as "R" in the connection table).

The program calculates all fragments obtain-

able by breaking bonds. There are no assumptions about probabilities for fragmentations or peak intensities; only the intensities of isotope peaks are relevant. The number and kind of bonds to break can be set by the user. For each generated fragment the program calculates the isotope signals that are expected from its elemental composition.

As a second step, the program asks for a mass spectrum, which can be loaded as an ASCII file or typed in manually. The file format is very simple, so that programs to convert data from the spectrometer to the requested format can be easily generated. The spectrum may be edited as text later on.

The program then correlates the mass of the calculated fragments with the peaks of the experimental spectrum. At this point, TIMS does not use any rules and matches only the masses or $m \pm 1$ in case that H-shifts are allowed. At this point, the computational part of TIMS is terminated and the calculated and experimental spectra are displayed graphically.

TIMS displays the results on six different windows; four of them are text windows and two are graphics. The text windows are only a documentation of the results in a tabulated form; the user can select text and copy it into the clipboard. Only the window containing the experimental data is fully editable.

The two graphic windows can be thought of as a graphic user interface, which offers many possibilities for the user to obtain information. In the window containing the structure, the cursor assumes the shape of a hammer. It can be used to break bonds manually just by clicking at them with the mouse (as shown in Fig. 5).

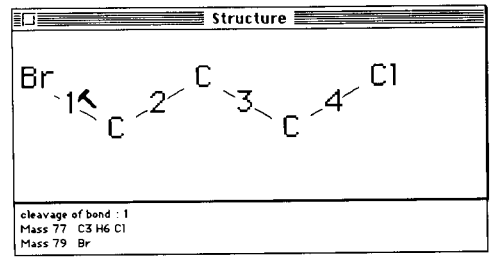


Fig. 5. Breaking bonds "by hand".

In the menu bar, the user can choose the combination of spectra he wants to see. The possibilities available are given in Table 1.

In the spectra display, the user can select the different peaks to obtain information about them. For experimental peaks the program calculates possible compositions, their double bond equivalents and, if requested, their high-resolution masses. For calculated fragments it delivers the composition and the bonds that have to be broken to obtain them. Further, the user can select a range in the mass spectrum by dragging the mouse and the program will display the length of that range (in mass units) in a window. Releasing the mouse button over a second peak causes the program to calculate all possible compositions of the mass difference in order to explain neutral losses.

In the Options menu, the user can decide to interpret the whole mass spectrum or only a part of it. Further, the element constraints (initially set by the molecular formula) are editable. In this way one can, for example, scan the spectrum for all bromine-containing fragments.

The number and kind of bonds to be broken can be changed (Fig. 6); however, the default

TABLE 1

Possible combinations of spectra displays

Top spectrum	Correlated experimental peaks	Not correlated experimental peaks	All calculated fragments	Not correlated calculated fragments
Bottom spectrum		All experimental peaks		Correlated calculated fragments

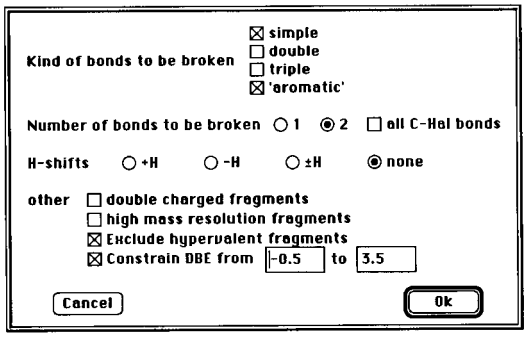


Fig. 6. Option dialogue.

values are useful in most instances. For rationalization of the possibly greatest number of peaks, the number of bonds should be 2; for distinguishing two possible structures with an experimental spectrum it is advisable to start breaking only one bond. Allowing H-shifts causes peaks to be correlated with fragments (or their isotope signals) with a mass difference of 1 and vice versa.

Enabling the option "double charged fragments", TIMS looks for peaks at $m/2z$. In this instance, low-resolution peaks with half mass units are obviously relevant. A further dialogue will ask the user when to correlate a peak at $m/2z$ with a fragment at m/z.

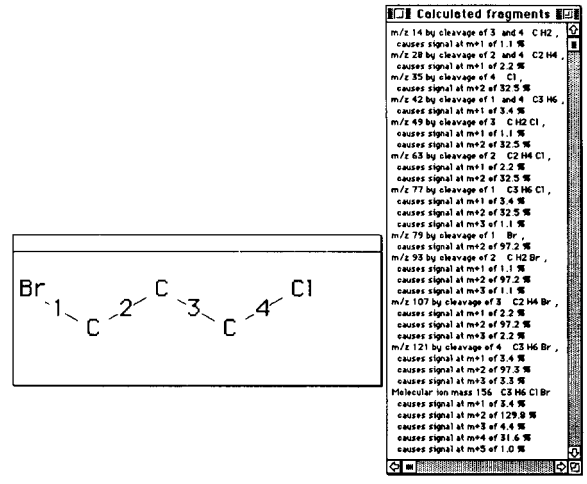


Fig. 7. Screen hard copy of structure and fragments windows.

"High resolution" means that TIMS will calculate the exact masses of the proposed compositions of experimental peaks when clicking on them in the graphics display. "Exclude hypervalent fragments" prevents the display of fragments with a highly improbable composition. "Constrain double bond equivalents" to a certain range will filter any species outsides the selected interval.

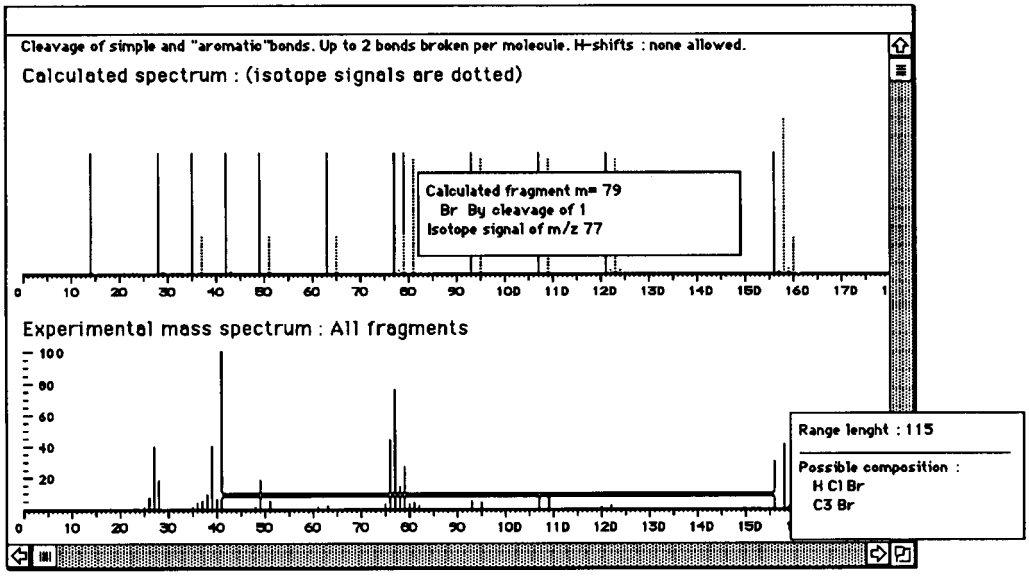


Fig. 8. Graphics display of calculated vs. experimental spectra.

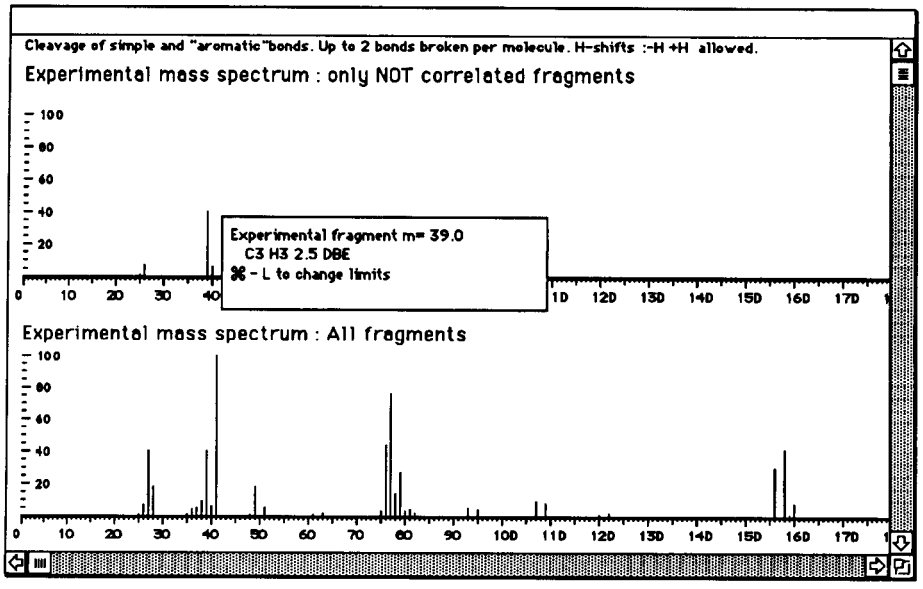


Fig. 9. Graphics display of experimental peaks.

Implementation and limitations. The file containing the data of the elements can easily be modified if the user wants to introduce new elements or isotopes such as ^{2}H and ^{14}C. TIMS can work with up to ten different kinds of elements simultaneously. This limitation should not be a serious restriction.

The greatest limitation of TIMS is due to the fact that it still works with static variables and cannot allocate more than 32 kB of memory. This causes limitation for the non-hydrogen elements to a maximum of 25 and for the number of possible fragments to 900. Efforts are being made to implement dynamic data storage. TIMS is System 7 compatible and supports Balloon Help.

An example. The most important features of TIMS are demonstrated for 1,3-bromochloropropane. Figure 7 shows the structure window and the list of calculated fragments.

Figure 8 shows the calculated and the experimental spectra. In the upper spectrum, the user clicked at the peak at m/z 79 and received information about it. In the lower spectrum, the user selected a range starting from the molecular ion to the base peak at m/z 41.

Figure 9 shows the comparison of the non-correlated experimental peaks with the whole experimental spectrum. The peak at m/z 39 that could not be explained corresponds to the cyclopropenyl ion.

DISCUSSION

SpecTool was developed as a tool to help the chemist in interpreting spectra. It supplies data (stored or calculated with the included tools) which support him in making decisions. However, SpecTool does not use "hidden heuristics", i.e., none of the programs use encoded knowledge. Hence it is not designed for individuals without some knowledge of spectrum interpretation. On the other hand, it can catalyse the routine work of chemists working in elucidating or verifying the structure of organic compounds. This system is not intended to replace spectroscopic databases, but at least it could eventually replace printed reference data collections. Currently the scope of the data and of the integrated programs is being increased.

This work was partly supported by the Swiss National Science Foundation.

REFERENCES

1 Q.N. Porter, Mass Spectrometry of Heterocyclic Compounds, Wiley, New York, 2nd edn., 1985.
2 G.R. Waller and O.C. Dermer (Eds.), Mass Spectrometry Reviews, Wiley, New York, 1982.
3 Mass Spectrometry Data Centre, The Eight Peak Index of Mass Spectra, Royal Society of Chemistry, Cambridge, 4th edn., 1991, 3 vols.
4 S. Heller, in J. Zupan (Ed.), Computer-Supported Spectroscopic Databases, Horwood, Chichester, 1986, p. 118.
5 K. Varmuza, Anal. Chim. Acta, (1993) in press.
6 NIST/EPA/MSDC Mass Spectral Database, National Institute of Science and Technology (NIST), Gaithersburg, MD.
7 Registry of Mass Spectra, Wiley, New York.
8 SpecInfo, Chemical Concepts, Weinheim.
9 S.E. Stein, P. Ausloos and S.G. Lias, J. Am. Soc. Mass Spectrom., 2 (1991) 441.
10 M. Farkas, M. Cadisch and E. Pretsch, in E.J. Karjalainen (Ed.), Scientific Computing and Automation (Europe), Elsevier, Amsterdam, 1990, pp. 455–466.
11 M. Cadish, M. Farkas, J.T. Clerc and E. Pretsch, J. Chem. Inf., 32 (1992) 286.
12 M. Cadisch, E. Pretsch, Fresenius' J. Anal. Chem., 344 (1992) 173.
13 K. Parsaye, M. Chingnell, S. Khoshafian and H. Wong, Intelligent Databases: Object Oriented, Deductive Hypermedia Technologies, Wiley, New York, 1989.
14 N. Woodhead, Hypertext and Hypermedia: Theory and Applications, Sigma Press, Wilmslow, 1990.
15 L. Bielawski, R. Lewand, Intelligent Systems Design: Expert Systems, Hypermedia and Database Technologies, Wiley, New York, 1991.
16 A. Fürst, J.T. Clerc and E. Pretsch, Chemometr. Intell. Lab. Syst., 5 (1989) 329.
17 H. Kubinyi, Anal. Chim. Acta, 247 (1991) 107.
18 E. Pretsch, J.T. Clerc, J. Seibl and W. Simon, Tables of Spectral Data for Structure Determination of Organic Compounds, Springer, Berlin Heidelberg, 3rd edn., 1986.
19 H. Kubinyi, in J. Gasteiger (Ed.), Software Entwicklung in der Chemie 2, Springer, Heidelberg, 1988, p. 167.
20 T. Brodmeier, Diplomarbeit, Laboratorium für Organische Chemie der ETH Zürich, Zurich, 1990.
21 M.M. Siegel and G. Gill, Anal. Chim. Acta, 237 (1990) 459.

Analytica Chimica Acta, 277 (1993) 305–311
Elsevier Science Publishers B.V., Amsterdam

Pattern recognition studies of tandem mass spectra

D. Swain and W.J. Dunn, III
University of Illinois at Chicago, Chicago, IL 60612 (USA)

R.E. Talaat
ABC Laboratories, 7200 E. ABC Lane, Columbia, MO 65205 (USA)

(Received 6th August 1992; revised manuscript received 20th November 1992)

Abstract

Principal components analysis and pattern recognition (PCAPR) techniques were applied to MS–MS spectra of fourteen organic compounds. Each spectrum was represented as a two-dimensional matrix containing information from the MS1 spectrum as well as from one, two or three MS2 spectra. The data were reduced by calculating a one-principal component model for each spectrum which explained between 86 and 99% of the variance. Each model was used to calculate each of the spectra, and residual standard deviations (R.S.D.s) were used as a measure of spectral similarity: low R.S.D.s (< 1.0) corresponding to similar spectra and higher R.S.D.s (> 1.0) to dissimilar spectra. The system shows promise for use in monitoring situations in that MS–MS spectra can be efficiently reduced and stored as principal components models and R.S.D. calculations can be used to identify a compound based on how well its spectrum is predicted by the available reference models.

Keywords: Mass spectrometry; Principal components analysis; Pattern recognition; Tandem MS

There are many methods available to the analytical chemist seeking to identify and quantify toxic or hazardous compounds in the environment. Every method has its limitations, however, and lately, increasing interest has been focussed on the hyphenated methods as a way to combine the attractive features of more than one technique [1–4]. Gas chromatography–mass spectrometry (GC–MS) in particular is currently the primary monitoring tool of the United States Environmental Protection Agency (EPA) [4].

Chromatography is often the lengthiest step in a hyphenated system, and it is sensible to try and eliminate that step by analyzing mixtures directly. This has been done with MS–MS [5] by using a soft chemical ionization in the first stage (MS1) to separate molecular ions; then the second stage (MS2) can be used to acquire fragmentation patterns of the separated molecular ions. In this application, MS1 can be likened to an instantaneous "chromatography column" that separates molecular ions on the basis of mass, and MS–MS can be used as a sort of "hurry up" GC–MS. However, information from a second stage of fragmentation is not obtained.

In the present work, principal components analysis and pattern recognition were applied to the reduction, storage and recognition of reference MS–MS spectra of pure compounds. We are using this approach to develop a general method for compound identification based on MS–MS spectra, and ultimately we hope to be able to use these techniques for the direct analysis of mixtures from MS–MS spectra that contain information from two stages of fragmentation.

Correspondence to: D. Swain, University of Illinois at Chicago, 833 S. Wood St., Chicago, IL 60612 (USA).

EXPERIMENTAL

Computations

All computations were carried out on a Zenith Z-386 personal computer equipped with a math coprocessor. The UNIPALS software for principal components analysis and partial least squares regression developed in this laboratory [6,7] was used for computations.

Data collection

The fourteen compounds included in this study are shown in Fig. 1. Spectra of 500-ng samples of 3,3′-dimethoxy benzidine (DMB), (2,4-dichlorophenoxy)acetic acid (24D) and (2-methyl-4-chlorophenoxy)acetic acid (MCPA) were collected under particle beam LC–MS–MS conditions. The samples were introduced onto a Spectra-Physics SP8700-XR liquid chromatograph coupled to an Extrel Thermabeam™ interface. The MS–MS spectra were collected on a Finnigan MAT 4500 Triple Stage Quadropole™ mass spectrometer interfaced with the Extrel Thermabeam™. The collision energy was 20–30 eV. Three MS2 spectra were collected for 33DMB, and two each for 24D and MCPA.

Spectra for approximately 1-μg samples of eleven aromatic hydrocarbon samples were collected on a VG TRIO-3 Triple Quadropole mass

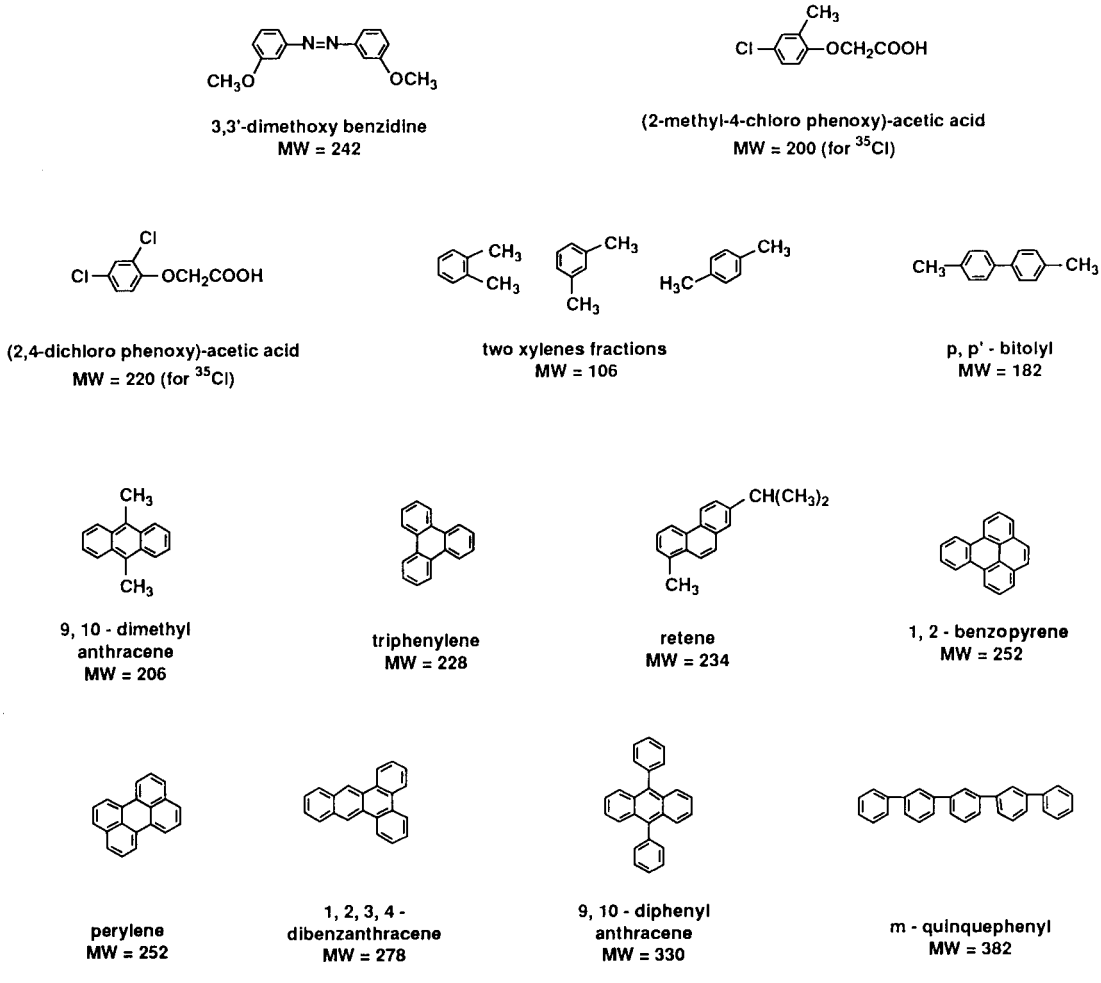


Fig. 1. The fourteen samples for which MS–MS spectra were obtained.

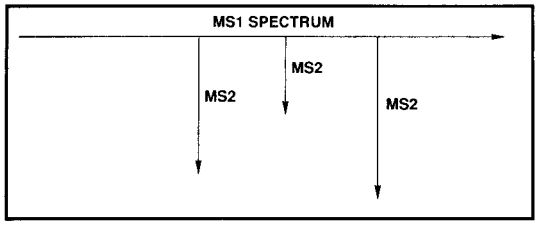


Fig. 2. Matrix representation of an MS–MS spectrum.

spectrometer interfaced with a Hewlett Packard Model 5890A gas chromatograph. The samples included two GC-separated fractions of the three isomeric xylenes (XY1 and XY2), *p,p'*-bitolyl (BIT), 9,10-dimethylanthracene (DMA), triphenylene (TRI), retene (RET), 1,2-benzopyrene (BP), perylene (PER), 1,2,3,4-dibenzanthracene (DBA), 9,10-diphenylanthracene (DPA) and *m*-quinquephenyl (QQP). The collision energy was 50 eV. A single MS2 spectrum (of the molecular ion) was collected for each sample.

Data preprocessing

Ion current intensities were arranged in a 36 × 345 matrix for each compound, shown schematically in Fig. 2. All intensities less than 1% of the base peak were entered as zeros; unavailable information was also entered as zeros.

The MS1 spectrum was represented in the first row, spanning mass-to-charge ratios of 40 to 384. Each MS2 spectrum was represented as a column emerging from its parent ion intensity. Row two corresponds to the sum of all MS2 intensities for loss of 1–10 from the parent; row three corresponds to loss of 11–20; etc. So the 35 additional rows after the MS1 spectrum include information on losses of up to 350. All MS2 intensities were scaled so that the parent ion intensity is a relative 100% in a given fragmentation column.

Principal components analysis

A principal components model for each spectrum was calculated according to Eqn. 1.

$$x_{ij} = X_j + \sum_{a=1}^{A} t_{ia} b_{aj} + e_{ij} \qquad (1)$$

Here i is a row (MS2 mass interval) index, j is a column (MS1 mass interval) index, and A is the number of principal components in the model. x_{ij} is the original matrix element, X_j is the average for column j, and e_{ij} is the model residual.

The percent variance explained by a given model is given by Eqn. 2.

$$\%\mathrm{VAREXP} = \sum_{j=1}^{345} \sum_{i=1}^{36} \left[(x_{ij} - X_j)^2 / \left(\sum_{a=1}^{A} t_{ia} b_{aj} - X_j \right)^2 \right] \times 100\% \qquad (2)$$

Pattern recognition

The reduced data for each sample can be visualized as a pattern of points in parameter space. Pattern recognition was carried out by calculating residual standard deviations (R.S.D.s) as a measure of the distance, or dissimilarity, between compounds. Each model was used to predict each spectrum, and Eqn. 3 gives the residuals.

$$(e_{ij})_{\mathrm{RS}} = (x_{ij} - X_j)_{\mathrm{R}} - \left(\sum_{a=1}^{A} t_{ia} b_{aj} \right)_{s} \qquad (3)$$

These residuals, shown here for the use of model "S" to predict spectrum "R", were then used in Eqn. 4 to get the corresponding R.S.D.

$$(\mathrm{RSD})_{\mathrm{RS}} = \sum_{j=1}^{345} \sum_{i=1}^{36} (e_{ij})^2_{\mathrm{RS}} / [(36 \times 345) - 2] \qquad (4)$$

RESULTS AND DISCUSSION

Data reduction

The increased quantities of information available from MS–MS, and from hyphenated techniques in general, can be a problem in monitoring situations because of the computer space required to store reference spectra. This is one of the reasons that reduction of the data by principal components analysis is so desirable. For the compounds studied and the preprocessing used, we were able to achieve data reduction [in kilobytes (kB) of computer storage space required]

TABLE 1

Percent variance explained by one- and two-principal component models of the fourteen 36×345 matrices

Compound	One-PC (%)	Two-PC (%)
DMB	95.7	98.3
24D	95.9	96.3
MCPA	89.0	95.7
XY1	96.4	100
XY2	95.9	100
BIT	93.0	100
DMA	96.5	100
TRI	99.1	100
RET	91.9	100
BP	99.4	100
PER	89.7	100
DBA	98.9	100
DPA	86.8	100
QQP	97.8	100

on the order of 90% (around 50–5 kB) for a one-component model, and on the order of 80–90% (around 50–7 kB) for a two-component model.

Percent variance explained

Table 1 shows percent variances explained for each sample with a one- and two-component model. The one-component models are quite good, and since they give us the greatest data reduction we used a simple one-component model in all further studies. The fact that the data preprocessing allows inclusion of full mass ranges in the generalized spectrum for both stages of MS, along with the fact that simple models of these matrices capture the data well, speak well for the mode of data input used in this study.

Residual standard deviations

Table 2 shows the residual standard deviations (R.S.D.) calculated when each of the fourteen one-component models is used to predict all fourteen original spectra. The diagonal elements correspond to models used to predict the spectra from which they were derived ("self") whereas off-diagonals correspond to cross-mappings ("non-self").

For identification purposes, there is good discrimination between "self" and "non-self". The diagonal elements of this table are all comfortably less than one (the highest is 0.55) whereas the nondiagonals are generally all greater than one. This is significant because it shows that if an unknown spectrum of one of these fourteen compounds were mapped against each of the fourteen models, an R.S.D. of less than 1.0 would be an accurate marker for compound identification.

The exceptions to this occur (predictably) where there is spectral similarity, as in the case of structural isomers. Virtually identical off-diagonals were calculated for the two xylene fractions studied. This lack of isomer discrimination by

TABLE 2

Residual standard deviations for each spectrum (horizontal) fit to each model (vertical)

	DMB	24D	MCPA	XY1	XY2	BIT	DMA	TRI	RET	BP	PER	DBA	DPA	QQP
DMB	*0.24*	1.62	1.68	1.28	1.28	1.27	1.32	1.20	1.56	1.24	1.41	1.16	1.58	1.07
24D	1.20	*0.32*	1.68	1.29	1.28	1.27	1.32	1.21	1.56	1.25	1.42	1.16	1.58	1.08
MCPA	1.21	1.62	*0.53*	1.29	1.28	1.29	1.33	1.22	1.57	1.26	1.43	1.18	1.59	1.09
XY1	1.19	1.62	1.66	*0.24*	*0.27*	1.27	1.31	1.21	1.56	1.25	1.41	1.16	1.58	1.07
XY2	1.19	1.62	1.66	*0.25*	*0.25*	1.27	1.31	1.21	1.56	1.25	1.41	1.16	1.58	1.07
BIT	1.20	1.62	1.68	1.29	1.28	*0.32*	1.32	1.21	1.56	1.25	1.42	1.16	1.58	1.08
DMA	1.19	1.62	1.68	1.28	1.27	1.27	*0.24*	1.21	1.56	1.25	1.41	1.16	1.58	1.07
TRI	1.19	1.62	1.68	1.28	1.28	1.27	1.31	*0.11*	1.56	1.24	1.41	1.16	1.58	1.07
RET	1.20	1.62	1.68	1.28	1.28	1.27	1.32	1.21	*0.43*	1.25	1.41	1.16	1.58	1.07
BP	1.19	1.62	1.68	1.28	1.28	1.27	1.31	1.20	1.56	*0.10*	*0.58*	1.16	1.51	1.07
PER	1.19	1.62	1.68	1.28	1.28	1.27	1.32	1.21	1.56	*0.38*	*0.44*	1.16	1.51	1.07
DBA	1.19	1.62	1.68	1.28	1.28	1.27	1.31	1.21	1.56	1.24	1.41	*0.12*	1.58	1.07
DPA	1.19	1.62	1.68	1.28	1.28	1.27	1.32	1.21	1.56	1.21	1.38	1.16	*0.55*	1.07
QQP	1.19	1.62	1.68	1.28	1.28	1.27	1.32	1.21	1.56	1.25	1.41	1.16	1.58	*0.32*

TABLE 3

MS1 spectra for 1,2-benzopyrene and perylene

Benzopyrene		Perylene		Difference	
Mass	Intensity	Mass	Intensity	Mass	Intensity
99	1.29	99	–	99	1.29
100	2.19	100	1.32	100	0.87
112	11.76	112	7.37	112	4.39
113	22.49	113	10.90	113	11.59
114	–	114	1.95	114	1.95
124	16.61	124	27.36	124	10.75
125	33.19	125	–	125	33.19
126	33.95	126	30.14	126	3.81
198	1.50	198	1.51	198	0.01
199	1.14	199	–	199	1.14
200	1.77	200	1.55	200	0.22
222	3.94	222	3.13	222	0.81
223	4.11	223	3.17	223	0.94
224	10.47	224	7.86	224	2.61
225	5.41	225	3.73	225	1.68
226	6.66	226	3.71	226	2.95
227	1.24	227	–	227	1.24
246	2.37	246	1.90	246	0.47
247	2.93	247	2.55	247	0.38
248	18.21	248	16.39	248	1.82
249	13.45	249	11.89	249	1.56
250	55.05	250	49.91	250	5.14
251	23.52	251	17.96	251	5.56
252	100.00	252	100.00	252	–
253	38.52	253	38.06	253	0.54
254	4.83	254	4.06	254	0.77

MS–MS may be inherent to the chemical system, but it may also be due to the conditions of data collection: impure samples (one of the fractions contains two of the three isomers), small sample size and only one MS2 fragmentation per spectrum. Work is in progress to collect better (more discriminatory) MS–MS spectra of the *three* separated xylenes with more parent ion fragmentations, and to see if that can lead to well-separated patterns in parameter space for reduced models of the three spectra.

Somewhat better separated are the other isomers in the data set: 1,2-benzopyrene and perylene which are both unsubstituted polycyclic aro-

TABLE 4

MS2 spectra for 1,2-benzopyrene and perylene

Benzopyrene		Perylene		Perylene (rescaled)	
Mass	Intensity	Mass	Intensity	Mass	Intensity
41.64	13.60	41.31	14.12	41	7.66
202.04	0.12	–	–	–	–
225.05	0.19	–	–	–	–
226.12	1.61	226.11	1.44	226	0.78
250.10	4.77	250.12	5.24	250	2.84
251.19	4.04	251.86	84.33	–	–
252.20	100.00	252.28	100.00	252	100.00

matics with a molecular weight of 252. From Table 2, it is possible to see that BP models itself very well (R.S.D. = 0.10) whereas PER models itself somewhat more poorly (R.S.D. = 0.44), and that BP models PER worse than it does itself (R.S.D. = 0.58) and PER models BP approximately the same as it does itself (R.S.D. = 0.38).

The difference between an R.S.D. of 0.10 and 0.58 is probably significant although information is not yet available as to how reproducible that difference is.

Mass lists from both MS1 and MS2 of BP and PER are shown in Tables 3 and 4. The MS1 spectra appear to be virtually superimposable as there is only one substantial peak (at $m/z = 125$) which is present in one spectrum but absent in the other.

The MS2 spectra are different at two points. The peak at 41.31 in PER rounds off to 41 and ends up in row 23 of the matrix, whereas the peak at 41.64 in BP rounds off to 42 and ends up in row 22 of the matrix. Also, since the peak at 251.86 in PER appears to be distinct in the line spectrum, it was treated as a 251 peak and it shows up in row 2 of the matrix.

The question of whether the differences in R.S.D. values for BP and PER mapping onto the BP model were due more to MS1 differences or MS2 differences was addressed as follows. All MS2 information was removed from the MS–MS matrices for both BP and PER, leaving just the first-row MS1 spectrum and 35 additional rows of zeros. A single tiny intensity element (0.005% of the base peak intensity) was included in the second row, below $m/z = 252$, in order to accommodate software requirements that there be a second dimension of variable data in the matrix. This was done for both BP and PER, and then reduced models were derived and R.S.D.s calculated.

As indicated in Fig. 3, an R.S.D. of less than 0.001 was calculated for self-mappings of both BP and PER. This was expected since a one-component model of virtually a single row of significant data should be a nearly "perfect" model (i.e., close to 100% variance explained, and self-mapping R.S.D.s close to zero).

BP modelled PER with an R.S.D. of 0.19 and PER modelled BP with an R.S.D. of 0.22. These numbers give some idea of the magnitude of deviation to expect from MS1 spectra alone that appear to be virtually superimposable.

In order to determine the effect of the $m/z =$ 251 peak in PER (Table 4) on the calculated R.S.D.s, that peak (intensity = 84.33%) was treated by rounding it off to 252, adding its intensity to the 100% value already at 252, and rescaling the entire MS2 spectrum by dividing through by 1.8433. With this rescaling, the MS2 spectrum of BP and PER look a lot more similar (see Table 4), and that appearance is borne out by the calculated R.S.D.s (see Fig. 3). The 0.10–0.58 distinction, between the mapping of BP and PER spectra onto the BP model, was reduced considerably, to a distinction between 0.10 and 0.19.

The overall conclusions to be drawn from this work are that the data preprocessing described here provides a matrix representation of MS–MS spectra that results in good one-component models, and that the separation in parameter space is good for nonisomeric compounds which means that PCAPR can provide a useful technique for the rapid screening of compounds.

Future directions include expanding the data

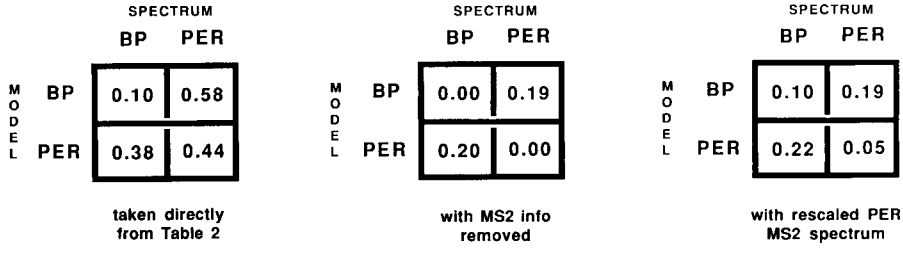


Fig. 3. Residual standard deviations for three different 1,2-benzopyrene–perylene mappings.

base, standardizing an acquisition technique, and exploring the reproducibility, or precision, of the data (i.e., how much R.S.D. should be expected for spectra from the same compound in independent measurements?). Once these goals are accomplished, true unknowns will be run through the standardized protocol in order to get an unbiased view of the predictive value of these techniques in screening situations.

The authors wish to thank Tammy Jones at the U.S. Environmental Protection Agency for supplying the MS–MS spectra for 33DMB, 24D and MCPA. The financial support of the University of Illinois at Chicago is gratefully acknowledged.

REFERENCES

1 T. Hirschfeld, Anal. Chem., 52 (1980) 297A.
2 C.L. Wilkins, Anal. Chem., 59 (1987) 571A.
3 D.F. Gurka, L.D. Betowski, T.L. Jones, S.M. Pyle, R. Titus, J.M. Ballard, Y. Tondeur and W. Neiderhurt, J. Chromatogr. Sci., 26 (1988) 301.
4 D.F. Gurka, L.D. Betowski, T.A. Hinners, E.M. Heithmar, R. Titus and J.M. Henshaw, Anal. Chem., 60 (1988) 454A.
5 R.W. Kondrat and R.G. Cooks, Anal. Chem., 50 (1978) 81A.
6 W.G. Glen, W.J. Dunn, III and D.R. Scott, Tetrahedron Comput. Methods, 2 (1989) 349.
7 W.G. Glen, M. Sarker, W.J. Dunn, III and D.R. Scott, Tetrahedron Comput. Methods, 2 (1989) 377.

Analytica Chimica Acta, 277 (1993) 313–324
Elsevier Science Publishers B.V., Amsterdam

HIPS, a hybrid self-adapting expert system for nuclear magnetic resonance spectrum interpretation using genetic algorithms

R. Wehrens, C. Lucasius, L. Buydens and G. Kateman

Laboratory of Analytical Chemistry, Catholic University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen (Netherlands)

(Received 3rd September 1992)

Abstract

An automatic system for the interpretation of two-dimensional NMR spectra of proteins, HIPS, is presented. Several artificial intelligence techniques are combined to form a flexible, hybrid system that has (limited) learning capabilities. Following the structure of the problem, the system is divided in modules with distinct functionalities. The first two modules are rule-based, and can be validated and refined semi-automatically using a set of already interpreted spectra. In this way, an optimized ruleset can be obtained to interpret unknown spectra. Results indicate a significant effect of training on performance. In the third module, a genetic algorithm is used to tackle a search problem of huge dimensions in which patterns found in the NMR spectra should be mapped to amino acids in the sequence.

Keywords: Nuclear magnetic resonance spectrometry; Expert systems; HIPS; Genetic algorithms; Hybrid expert systems

In recent years, much work has been done to make the determination of the three-dimensional structure of proteins possible. A technique that is most useful is two-dimensional nuclear magnetic resonance (NMR) spectrometry. A complete interpretation of the NMR spectra of a protein yields a set of distance constraints that can be used to calculate the conformation of the protein. The bottleneck, however, lies in the interpretation stage. This is a complicated process of trial and error that may take months or even years of expert's time. Automation of the process is clearly desirable.

Few programs have been written that cover the complete interpretation of two-dimensional protein spectra (e.g., [1–3]); other programs (e.g., [4–6]) cover parts of the process. Most of these programs depend on human input during the problem solving stage to keep the interpretation process manageable. Since experts in most cases have access to more detailed information than the computer programs, for instance regarding peak shapes and spin patterns, they are able to prune the number of possibilities significantly. This makes the task easier for the computer program. However, it is difficult to compare different systems in literature. In many cases, different types of input spectra are used, or spectra that have been simulated in different ways.

The use of expert systems for spectrum interpretation has several advantages. One of the most important advantages is that the chemical knowledge is encoded explicitly in the system, in a form that is natural to an expert. In conventional pro-

Correspondence to: R. Wehrens, Laboratory of Analytical Chemistry, Catholic University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen (Netherlands).

grams the knowledge is typically hidden in the code, which can make it difficult to communicate with the program during the interpretation, and to adapt the program if conditions change. Expert systems are more flexible and therefore more useful in an iterative interpretation process, since the expert using the system can easily understand the individual pieces of knowledge and use them for his or her own purposes. There are, however, problems that cannot be handled well except by brute force. Large-scale search, for instance, is a problem solving paradigma that is not easily implemented in purely rule-based systems. Therefore, hybrid systems, combining classical expert system techniques with other paradigms like neural networks or genetic algorithms can often be very useful.

In this paper, we will describe the interpretation of two-dimensional NMR spectra of proteins by such a hybrid expert system. Starting from the peak positions in several types of NMR spectra (see below), HIPS (Heuristic Interpretation of Protein Spectra) combines adaptive expert system techniques and genetic algorithms to automatically produce an ordered list of spin patterns matching the sequence of amino acids in the protein. This means that all protons near the backbone of the protein have been assigned to resonance positions in the spectra. With this information distance constraints can be set up, from which the overall three-dimensional structure of the protein can be calculated. In the following section, we will discuss the benefits of hybrid systems and some of the techniques used will be described. After that, we will give a more detailed account of the NMR interpretation process. Finally, interpretation results of three test proteins will be given.

HYBRID EXPERT SYSTEMS

Hybrid expert systems are systems that do not rely solely on knowledge-based inference, or more popularly said, if–then rules, to find solutions to problems. Also other techniques are applied if appropriate. Combinations of expert systems with the pattern recognition capabilities of neural networks [7,8] are the most common, but also other techniques like genetic algorithms [9] can be very useful. The latter technique exploits the structure in a solution space to perform a very efficient search (vide infra). The combination of classical expert system techniques with other paradigms allows for a very natural representation, in which each problem solver is used for its own type of problems. This leads to better performance and more flexible systems. However, relatively few hybrid systems have been built so far because of the difficulties encountered in coupling two or more sophisticated problem solving paradigms.

In this paper, a system is described that combines a self-adapting expert system with a genetic algorithm. The former is used for synthesis and pattern recognition subproblems, whereas the latter is used to sample a huge subspace of possible solutions. In the next sections, the two techniques will be discussed in greater detail.

Self-adapting expert system techniques

In the past, many expert systems have been built that, once optimized and delivered, were completely rigid. In many cases, it was extremely difficult even for the knowledge engineers who built the system to adapt the system to fit new requirements, and the inflexibility of such systems decreased their level of acceptance. Apart from the inflexibility issue, optimization of rule sets and validation of prototype expert systems are difficult tasks in themselves, for which no general strategy has been accepted yet. One of the approaches addressing these issues is called the refinement strategy [10,11]. It assumes that the knowledge in the knowledge base is essentially correct and needs only minor changes to obtain optimal performance. Whereas this assumption is a demanding one, in many cases it will be at least partially fulfilled. A set of solved cases is used to identify weak spots in the problem solving process, and the rules that are responsible for the flaws in the reasoning are identified on the basis of some simple statistics. Then, appropriate actions are considered and ranked to their expected merits. In the original system, SEEK [10], the user was presented the ordered list of alternatives and could select the one that appeared best. In

this way, the proposed changes would be validated by the user (in most cases the expert) selecting them. The approach has been automated in the successor of SEEK, SEEK2 [11], where the best alternative was chosen automatically. More sophisticated techniques, such as an extensive meta-language, were used to rank the different refinements to their merits. Although the SEEK and SEEK2 programs are very specific, the ideas behind the refinement approach are quite simple and therefore easily applicable in other domains, such as the present one.

The advantages of the refinement strategy are clear: not only is the system validated against a dataset of known cases, but also the parts of the system that perform inferiorly are identified and corrected if possible. Even if the requirement that the knowledge in the knowledge base is essentially correct is not completely satisfied, the performance may be improved. Furthermore, the user will be able to see where the refinement strategy fails in such a case and therefore it is easier to find larger errors. An additional advantage is that the refinement module in the expert system can be built as a separate subsystem, validating the results of the real system. This also leaves open the possibility to disconnect the refinement module when the system has been optimized. Including such a semi-automatic refinement module in an expert system will also increase the flexibility of the system. If new conditions are met (for instance a better analysis method), the refinement module is connected to the system that in this way can be trained with examples from the new situation. New optimal settings can be found without extensive reprogramming. The refinements apply only to rule-based (sub)systems. If more knowledge representations are used, other techniques should be used to obtain the same flexibility.

In the present application, refinements are limited to a number of cases in which a set of ordered alternatives is set up on the basis of supports that have been collected. As well the number of supports needed for acceptance as the relative importance of the individual supports could be refined to obtain the best set of alternatives. By restricting the refinements to these parameters, it proved possible to build a fast and strongly focussed refinement module that can be coupled to an existing expert system that should be refined, and decoupled from the system without leaving any further traces. In most cases the generality in the original SEEK systems is not needed, and the refinements can be limited to a small set of operators.

The refinements provide an easy way to adapt the system to external parameters. If, for instance, it is known that the protein of interest resembles the structure of another, already interpreted protein, then the latter can be used to tailor the expert system settings to the particular problem. One can even think of the possibility to train the expert system on partial results of the protein of interest. Another situation in which the flexibility of the system is useful, is the case in which not all spectra have been recorded. For instance, not all spectra are crucial for the interpretation and some may therefore be omitted. However, this influences the validity of the rules. New settings may be found then by training on other proteins without using these spectra.

Genetic algorithms

Genetic algorithms [9] form a class of problem solvers that are especially powerful in sampling a large solution space with many local optima. They usually start with a random population of candidate solutions, where each solution is represented as a string. Each string is evaluated according to a fitness criterion. The true solution is expected to have the highest fitness. In the next step, strings are allowed to reproduce themselves with a probability relative to their fitness. A crossover operator is used to combine parts of highly fit parent solution strings in order to obtain even better child strings. This way, the new population is largely derived from successful solutions. A small amount of random mutations prohibits the system to converge prematurely. The ultimate goal is to approach the ideal solution by combining those parts of successful trial solutions that contribute positively to the fitness.

The difficult part in using genetic algorithms lies in the fact that a convenient and efficient representation must be found, along with opera-

tors for crossover and mutation that match the representation. In numerical optimization problems, this is usually not so much of a problem, but in sequencing and subset selection problems, as in the present case, other operators than usual are necessary to obtain good results [12]. Furthermore, a fitness criterion must be defined that also matches this representation.

NUCLEAR MAGNETIC RESONANCE OF PROTEINS

In this section, the principles of NMR as applied in the structure determination of proteins will be reviewed briefly. More information can be found in the references and NMR literature.

In proton NMR, each peak in the spectrum can be assigned to a proton or an interaction between two protons. The position of each peak is determined by the chemical environment of the proton(s) related to it. Because the number of protons in proteins is very large, the spectra are very complicated. One-dimensional spectra can not be interpreted for this reason. Therefore, two-dimensional spectra are used and sometimes even spectra with more dimensions. However, recording three- and higher-dimensional spectra of sufficient resolution may take days of expensive machine time, and therefore it is worthwhile to try to interpret less complicated spectra that contain less information. In a two-dimensional spectrum the ordinary one-dimensional spectrum is more or less present on the diagonal, in principle showing a signal for each proton in the chemical structure. The off-diagonal crosspeaks give, dependent on the kind of spectrum, additional information. An example of the kind of spectra that is used is the NOESY (Nuclear Overhauser Enhancement Spectroscopy) spectrum [13]; in this kind of spectrum a crosspeak appears between the resonance positions of two protons if they are less than 5 Å apart. From a complete interpretation of a NOESY spectrum, where each proton is mapped to a peak position on the diagonal and each off-diagonal crosspeak indicates a through-space interaction of two protons, a set of distance constraints can be calculated. These can be used to construct a three-dimensional model of the protein, e.g., using the distance constraints algorithm [14]. The difficult part is to map the peak positions in the spectrum to protons in the chemical structure. For this, several types of spectra are used that give complementary information. The key to the interpretation is the fact that protons that have more or less the same chemical environment, give rise to peaks that are close together in the spectrum. For instance, most protons in the backbone of the protein that are attached to the nitrogen lie in an area between 6 and 10 ppm. In general, the expert may use information from peak positions, but also information about peak shapes and peak intensities to interpret the spectra.

Interpretation of protein spectra

The common approach is the "sequential assignment" strategy [13]. It consists of three stages. In the first stage, sets of peaks are identified that belong to one single amino acid in the protein. Such a set of peaks will hereafter be denoted as "pattern". It consists of a number of diagonal peaks, representing the protons in the amino acid, and a set of off-diagonal peaks, representing the interactions between these protons. As an example, two amino acids and their patterns are depicted in Fig. 1. Normally, only protons up to the γ position are included in the pattern, since their peak positions provide enough information to determine the structure of the protein backbone. The result of the patternsearch step consists of a number of patterns that eventually should be mapped to separate amino acids in the sequence.

Each pattern has distinct characteristics, depending on the type of amino acid it belongs to. Among these characteristics are for instance the resonance position of the C_β protons, the presence or absence of C_β or C_γ protons, and many others. In the second step of the spectrum interpretation, each pattern found in the first part is classified as an amino acid, sometimes a specific type like glycine, sometimes a group of amino acids, like aromatic amino acids. After this step, for each pattern a preliminary classification or a set of possible classifications is obtained.

In the third and last step, data from NOESY spectra are used, together with the results of the previous assignment part, to establish a list of all patterns that are possibly neighbours in the sequence. From this list, a sequence of patterns (or possibly more than one valid sequence) is obtained that maps onto the sequence of amino acids. At this point, all peaks present in the patterns are explicitly assigned to a proton in the protein. The assignment of the other protons is now relatively straightforward because of the large number of crosspeaks that have been assigned already. The construction of a valid sequence of patterns, however, is a large problem. To place seventy patterns (a medium-sized protein) in a specific sequence, there are $70! \simeq 10^{100}$ possibilities. In most cases, more patterns will be found than the number of amino acids, so the problem becomes even bigger. This results in a scenario where only the obvious patterns are interpreted in the first place (for instance, there is only one plausible pattern for the one tryptophan residue in the sequence), and where these assignments are used to further refine other assignments. This process continues iteratively until a satisfactory solution has been obtained.

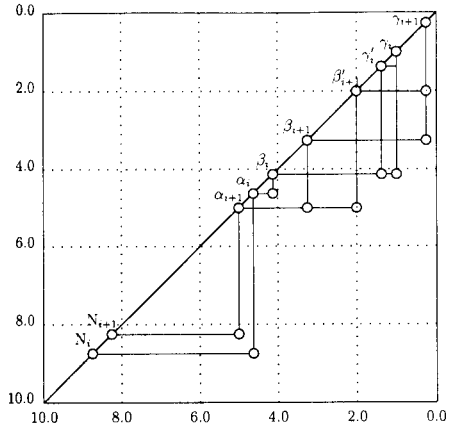


Fig. 1. Two amino acids and their schematic NMR patterns. Greek letters refer to the carbons to which a proton is bound.

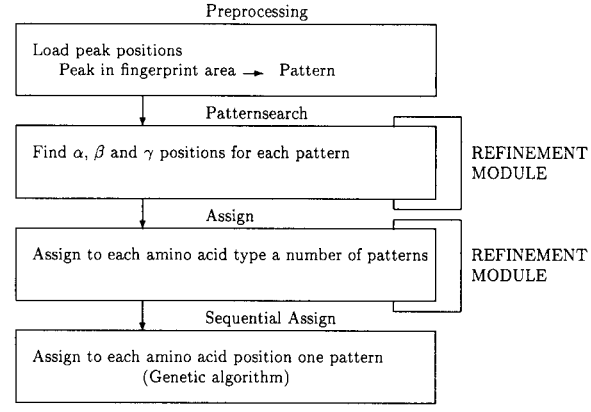


Fig. 2. The structure of HIPS. To the first two modules after the prepocessing stage, a refinement module has been attached to validate and optimize performance. This module can be disconnected when the system has been optimized. The last module incorporates a genetic algorithm to search for the best sequence of patterns that matches the sequence of amino acids.

IMPLEMENTATION AND RESULTS

In this section, the results of the expert system on the interpretation of three proteins will be discussed. For each of the three steps in the spectrum interpretation, pattern search, pattern assignment, and sequential pattern assignment, a separate module has been implemented (see Fig. 2). At each point in the interpretation, results can be inspected, saved, altered, and results of earlier sessions may be read. It should be noted that the system only uses peak positions with an uncertainty of 0.01 ppm, whereas an expert can also use information on peak shapes and intensities. Incorporating results of manual assignments will improve the performance of the system significantly. However, all results reported here are obtained without manual help, unless indicated otherwise.

The spectra of the proteins that have been used were obtained from the Department of Biophysical Chemistry at the University of Nijmegen.

TABLE 1

Characteristics of the spectra of the three test proteins
[The second column indicates the number of amino acids for which startpeaks are given in the literature (see text). The third column contains the number of peaks in the bottom half of the seven spectra for each protein]

Protein	Amino acids	Startpeak patterns	Peaks
E-L30	58	52	388
BPTI	58	52	953
Tendamistat	74	71	2868

They have been used as test spectra for the program PROSPECT [3] earlier. Seven types of spectra are used: NOESY, COSY and RCT spectra, recorded both in H_2O and D_2O, and DQSY spectra [13]. The spectra of two of the three proteins, BPTI (bovine pancreatic trypsine inhibitor) and Tendamistat, were constructed by taking the peak positions published in the literature [15,16] and simulating the NOESY spectra from the (known) three-dimensional structure. This way, the simulated spectra can be said to be very much like real spectra. Peaks may be missing, overlapping or unexpectedly be present in the same way as is found in experimental data. The third protein, E-L30, was measured at the laboratory of Biophysical Chemistry at the University of Nijmegen [17,18] and the DQSY spectra have been added manually. Some characteristics of the proteins and their spectra have been gathered in Table 1.

Patternsearch

In the first module of the expert system, a list of so-called start peaks is set up. These off-diagonal peaks lie in the fingerprint region in the spectrum where peaks are expected that are caused by interactions between the C_α and the amide protons. Each startpeak is considered to be part of a separate pattern. At least as many startpeaks as amino acids are expected [a]. From each startpeak, probable C_β positions are found using a number of criteria. If, for example, crosspeaks are found between an amide, C_α, and C_β resonance positions in as well NOESY as COSY spectra, then all three probably belong to the same amino acid. Several other crosspeaks can increase the faith in such a conclusion. It is obvious that not all crosspeaks are equally reliable or important, and therefore they are divided in three classes: minor, medium and major importance. Combinations of supports are handled explicitly in the rules. The same strategy is used in the search for C_γ peaks starting from the C_β positions found earlier.

The importance of each of the supporting crosspeaks is used to identify patterns in the spectra of the three test proteins. These are compared with the real patterns that should be found. A refinement module similar to the one originally used in SEEK [10] is used to fine-tune the importances of the individual crosspeaks and the amount of supporting evidence needed for a peak to be accepted as, for instance, a C_β peak. As an example, consider the situation where a peak is incorrectly accepted for the C_β position because of two supporting crosspeaks, N-β and α-2β. Then the following refinements may help to correct the error:

(i) diminishing the importance of the N-β crosspeak;

(ii) diminishing the importance of the α-2β crosspeak; and

(iii) raising the amount of evidence needed for acceptance above the total support for the incorrect peak.

Eventually, the refinements that will correct the largest number of errors are selected and presented to the user, who then can choose what refinement should be applied. This refinement process has been elaborated upon previously [19]. Since then, small changes have increased the performance to the level indicated in Table 2. The three proteins all have slightly different optimal settings, and a compromise has to be found to obtain the globally optimal settings. It ap-

[a] Exceptions are the amino acids proline, which does not have an amide proton, and glycine, which has two. In the former case, dummy patterns containing no peak positions are used, that can be filled in after manual interpretation; in the latter case, both startpeaks are taken into account, as well as the combination of startpeaks. Each glycine thus results in three separate patterns.

TABLE 2

Patternsearch results for the three proteins
(The third, fourth and fifth columns indicate the number of the patterns found correctly using the initial, optimal, and global settings, respectively)

Protein	patterns	OK (init.)	OK (opt.)	OK (global)
E-L30	52	41	43	43
BPTI	52	43	50	40
Tendamistat	71	40	65	65
Total	175	124	158	148

peared that the results of E-L30 were not very sensitive to the refinements that were applied; some peaks were missing from the spectra, and some patterns overlapped in such a way that the system could not conclude something else that it already did. This was different in the proteins BPTI and Tendamistat. There, and especially in the search for C_γ positions, the results were extremely sensitive to the refinements. The optimal settings for both proteins were rather different: BPTI required less strict acceptance criteria than Tendamistat. This is clearly visible in the large number of errors in BPTI in the globally optimal settings whereas the optimal BPTI limits yield only two errors. However, errors in the C_γ positions are less important than errors in the C_β positions, and it was felt that the global performance level was satisfactory.

Assignments of patterns to types of amino acids

In the second module of HIPS, the patterns that have been found in the first part are assigned to types of amino acids. For this, characteristics of the patterns are used. These characteristics include the presence or absence of certain peaks, and the resonance positions of the peaks. If, for instance, a C_α and a C_β share crosspeaks in the aromatic region, then it is probable that the pattern belongs to one of the aromatic amino acids (phenylalanine, tryptophan, tyrosine and histidine). These crosspeaks then are considered in greater detail to be able to distinguish between the separate aromatic amino acids. Furthermore, specific regions in the spectra give information about the peaks that are in them, such as the aromatic region mentioned previously [20]. Thus, by looking at the positions of the peaks in a pattern one is able to predict the amino acid type, or at least a set of probable types. This is implemented in rules, where the presence or absence of each characteristic contributes for an overall certainty factor for a type of amino acid. A low certainty factor of a pattern for a specific amino acid type indicates that there are no or few indicators that the pattern belongs to that type; a very high certainty factor indicates that the pattern is most probably of that type. Altogether, approximately one hundred contributions to certainty factors are possible. This large number makes it impractical to use classes of supports like minor, medium and major, as in the patternsearch module. Instead, numbers are used. No explicit boundaries are given, but initially all contributions were in the interval between -30 and $+80$. They are combined by summing. After all contributions have been gathered, the best patterns are selected for each type of amino acid. If for instance five amino acids of type threonine are present in the sequence, then the five patterns that have the highest certainty factor for threonine are selected, and all others that have the same certainty factor for threonine. This way, the chances are very small so that a pattern belonging to for instance a threonine is classified otherwise.

However, patterns may be classified in more than one class of amino acids. It is very important that the correct amino acid is among the possibilities, because otherwise it is very difficult in the last module, the sequential assignment, to provide a correct sequence of patterns. In this case, so-called false negative (FN) errors, where the correct answer is not in the list of possible solutions, are far more serious than false positive (FP) errors, where more patterns, including the correct ones, are possible for a specific type of amino acids. The first aim of the refinement module is therefore to minimize the number of false negative errors of the ruleset. However, a large number of false positive errors will lead in the sequential assignment to many solutions that seem equally probable. A second objective is

therefore to maximize the discriminating abilities of the expert system by minimizing the number of false positive errors. Between these two objectives, an equilibrium should be found.

In Table 3 the results of refining the contributions to the certainty factors are given in a similar way as in the first module, the patternsearch. The results are based on the patterns that were found using the globally optimal settings of the patternsearch module. In the case of BPTI, for example, twelve patterns were incorrect in the patternsearch part. As will be clear, in many cases this does not prevent the assignment module to correctly classify most of them. In the second column, the number of patterns is given for which the correct amino acid type was concluded using the initial settings (which were obtained from an expert). Refinements, in which the number of incorrect assignments was minimized, yielded settings that gave the results in the fourth column in Table 3. In all cases a very good performance was achieved. The third and fifth columns contain the numbers of false positive errors of the initial and optimal settings, respectively. As well regarding the number of false negative errors as regarding the false positive errors, remarkable improvements can be achieved. From the optimal settings of the three proteins, a set of globally optimal settings was derived. The results obtained with these settings are also included in the table. In approximately fifty percent of the cases, the patterns that were assigned incorrectly to an amino acid were already incorrect in the patternsearch module. However, as can be seen in the BPTI results, an incorrect pattern can be classified correcly, especially if errors are at the C_γ position.

Sequential assignment of patterns

After the patterns have been assigned to types of amino acids, only one step is needed to have a complete interpretation of the NMR spectra for the backbone, C_β and C_γ protons. This step comprises a specific mapping of separate patterns to amino acids in the known sequence. For this, the information derived in the second module is used to ensure that each pattern matches the amino acid at its position in the sequence. In addition, peaks connecting neighbouring patterns that can be found in the NOESY spectra are used to determine whether two patterns may be neighbours in the chain. This way, for each pair of amino acids that occurs in the sequence, a look-up table of possible pattern pairs can be set up. Two criteria for a valid pair of patterns are applied:

(i) both patterns should match the types of the amino acids in the pair; and

(ii) Both patterns should be connected by crosspeaks in the NOESY spectrum. These connections can be crosspeaks between the amide-position of the second pattern and the amide, C_α or C_β positions of the first pattern.

Several pruning procedures, controlled by flags, can be used to limit the number of possibilities:

(i) All patterns that are used to construct a glycine pattern with two startpeaks and themselves only contain one startpeak are excluded from the sequential analysis. These patterns are

TABLE 3

Assignment results for the three proteins

(All results have been obtained with the global settings from the patternsearch module. OK indicates the number of correctly interpreted cases, FP indicates the number of fals positive cases)

Protein	OK (init.)	FP	OK (opt.)	FP	OK (global)	FP
E-L30	32/52	147	45/52	191	41/52	191
BPTI	46/52	185	49/52	182	47/52	217
Tendamistat	58/71	208	61/71	133	62/71	259
Total	136/175	540	155/175	506	159/175	667

frequently assigned to a glycine pattern, but except in the case of extreme overlap, this is incorrect. This step is actually already performed in the second module, since otherwise too few glycine patterns may be selected.

(ii) All combinations of patterns that can not be matched with eachother are excluded. If, for instance, for a combination of amino acids AA-1 AA-2 only the combination PATTERN-1 PATTERN-2 is possible, all pattern combinations for any other amino acid combination that begins with AA-2 are forced to start with PATTERN-2. Any other pattern combinations are deleted. If AA-2 occurs more than once, this is of course taken into account.

(iii) All patterns that have not been assigned to any amino acid type will be excluded from the sequential assignment step. This, in general, will not cause any correct patterns to be removed. In both the cases of BPTI and Tendamistat, two patterns were discarded that should have been included in the sequence. In all four cases, however, the patterns were misinterpreted in the patternsearch module, and only their amide and C_α positions were correct. Their places in the sequence are taken by other patterns.

This way, the number of possibilities is reduced as much as possible. In Table 4 the number of different amino acid pairs and their associated number of pattern pairs is given for the three proteins, after pruning. For each protein the optimal settings (in both the patternsearch and the assignment modules) as well as the global settings are used. Although the E-L30 protein is of the same size as BPTI, the number of possible pattern pairs is significantly larger in the latter case if global settings are used. This is a logical consequence of the composition of the proteins: some amino acids are more easily recognized than others, and therefore give rise to fewer possibilities in the pattern combinations. Tendamistat, again, poses the largest demands on the system.

The solution space for this problem is huge. For example, the number of solutions is $97!/(97-74)! \simeq 10^{130}$ for Tendamistat using the global settings. Some approaches break down the problem in smaller subproblems [3] by only tackling

TABLE 4

Number of different amino acid pairs in the three test proteins, and number of pattern pairs matching the amino acid pairs

[An amino acid pair may be present more than once in a protein. The numbers in the table are obtained after pruning (see text). In the cases of BPTI and Tendamistat, the optimal settings give a smaller number of pattern combinations than the global settings. In the case of E-L30 this effect is not observed because of the relatively large number of false negative errors in the assignment module using the global settings. In BPTI (manual), a number of incorrect pattern pairs has been eliminated by hand so that for each combination of amino acids not more than ten pattern combinations are possible]

Protein	Amino acids	Amino acid pairs	Patterns	Pattern pairs
E-L30 (opt.)	58	52	61	1032
E-L30 (glob.)	58	52	61	922
BPTI (opt.)	58	54	63	1032
BPTI (glob.)	58	54	64	1567
BPTI (manual)	58	54	63	352
Tendamistat (opt.)	74	64	90	2244
Tendamistat (glob.)	74	64	97	4042

part of the amino acid sequence at once, but the resulting overall sequence does not have to be the best one in that case. Also, such approaches rely heavily on human intervention. For instance, in the case of the semi-automatic assignment of BPTI with the program PROSPECT [3], 80% of the patterns was unambiguously assigned to one type of amino acid. In our automatic assignment module, not more than 20% of all patterns, including glycines and prolines was assigned to only one type of amino acids, to avoid incorrect assignments. Other programs require the number of input patterns to be equal to the number of positions in the sequence [4]. Furthermore, in our case we deal with realistic NOESY simulations that may yield a large number of sequential connections between patterns. In other programs, the number of sequential connections is much smaller, either because smaller proteins are used [2] or because less NOESY peaks are included in the input files [1].

As already said, a genetic algorithm is a search technique for large solution spaces with many local optima, and therefore can be used in an

automatic spectrum interpretation program. The initial population of solutions is constructed by making random permutation strings of patterns with the length of the sequence of amino acid combinations. The evaluation function is very simple: the pattern combinations that satisfy the two criteria concerning sequential crosspeaks and pattern assignments are counted. Thus, the highest fitness that can be obtained is $N-1$, where N is the number of amino acids in the sequence. In practice, this will not be achieved, since errors in the earlier interpretation stages may cause a "correct" pattern combination to be absent in the look-up table. However, if the previous assignment modules are working correctly, then the fitness of the true solution should be close to the maximally obtainable fitness. Other deviations are possible because of missing sequential connectivities, and patterns that have not been identified because their startpeaks are missing from the fingerprint area. In the latter case, the place of such a pattern will be taken by another.

In Fig. 3 the fitnesses of a typical genetic algorithm run are plotted on the y-axis. The algorithm converges very fast to a point where the fitness stabilizes somewhat and then by chance finds a permutation that allows for further improvement. As can be seen in the second column of Table 5, the fitnesses of the true solutions of

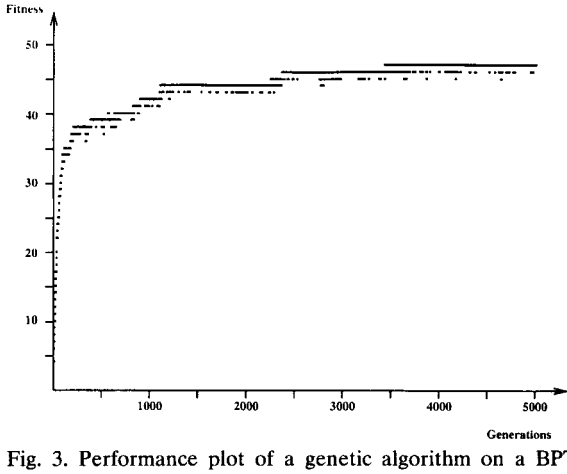


Fig. 3. Performance plot of a genetic algorithm on a BPTI dataset obtained with optimal settings. As the number of generations grows, the fitness increases to a level where no improvement can be found.

TABLE 5

Results of the sequential assignments with the genetic algorithm

(The fitness of the "true" solution is given in the second column. For each protein, five runs were done. The mean fitness obtained in these runs for each protein is given in the third column. If a pattern was placed in at least two of the five runs at the same position, it was considered to be a definite assignment. The number of assignments and the number of correct assignments for each protein are gathered in the last columns)

Protein	True F	Mean F	Assign-ments	Correct
E-L30 (opt.)	29	33.8	40	15
E-L30 (glob.)	27	34.2	37	17
BPTI (opt.)	42	40.4	46	30
BPTI (glob.)	37	43.8	40	18
BPTI (manual)	42	42.8	53	41
Tendamistat (opt.)	54	61.0	58	17
Tendamistat (glob.)	55	62.4	47	12

the three proteins lie below the maximally obtainable fitness $N-1$. The difference is most distinctly present in the case of E-L30, which also proved to be the most difficult case in the previous interpretation stages (see Tables 2 and 3). Remarkable is also the difference in the fitnesses of the true solutions in the case of BPTI, where the optimal settings of the assignment and patternsearch parts clearly provide a better result for the fitness of the true sequence.

In a real-world situation, however, one does not have the true solution, and the highest fitness obtained in a number of runs may be taken as the fitness of the "true" solution. However, given the stochastic nature of the genetic algorithm, it is better to combine results of a number of runs to form a global solution. In Table 5, results of five runs for each protein, both with global and optimal settings in the previous interpretation parts, are gathered. In each case, a population of 500 candidate solutions was used. For E-L30 and BPTI, 3000 generations, and in the case of Tendamistat, 5000 generations were taken. In the cases of BPTI and E-L30, fitnesses as high as the fitness of the correct solution were found in almost all cases within 200 generations. For Tendami, not more than 1000 generations were needed in general. The fitness obtained with the

global settings are in general higher than the fitnesses obtained with the optimal settings for each protein. This is as expected since more possibilities exist to construct a valid sequence in the case of the global settings (see Table 4).

A pattern is assigned to a position if it is at least in two runs present in the same position. This way, 66–88% of the positions were assigned patterns. In all cases, the amino acid on the assigned position matched the assignment of the pattern. However, only 26–65% of the assignments was in agreement with the "correct" assignments. It must be remembered, however, that this is not the result of the genetic algorithm failing to find the global optimum, but rather the genetic algorithm finding solutions that are in better agreement with its input data than the "true" solution. This is also the reason why the fitnesses found by the genetic algorithm are generally higher than the fitness of the "true" solution. Striking is the fact that in protein E-L30 more patterns are assigned correctly using the global settings, than are assigned correctly with the optimal settings. This may be a consequence of the small number of runs that is done with the genetic algorithm, and in which a pattern is already assigned to a position if it occurs twice.

To investigate the ability of the genetic algorithm to produce valid results starting from a good look-up table, the dataset provided by the global settings in the case of BPTI was pruned manually, so that each combination of amino acids contained maximally ten combinations of patterns. This way, the total amount of pattern combinations was diminished to 352, one-third of the original amount. The results of the runs with this input set have also been gathered in Table 5. In this case, over 77% of all assignments was in agreement with the true sequence.

The above results prove the usefulness of genetic algorithms to solve the problems encountered in the sequential assignment of protein spectra. However, the complete automatic assignment is not yet feasible, and manual input would greatly improve the results of the system. This is clear from the results with the manually pruned BPTI dataset. However, as already said, in real life the expert can at any point intervene and focus the interpretation results. Therefore, the pruned dataset constitutes a reasonable simulation of reality.

HARDWARE AND SOFTWARE

The expert system is written in KEE (Intellicorp Inc.), version 4.0. The six knowledge bases take approximately 150 K. Some 200 K of compiled LISP code is used. For patternsearch and assignments, not more than 5 min real-time is required for Tendamistat on a SUN SPARC-1 workstation. BPTI and E-L30 are done more than twice as fast. The genetic algorithm is written using the GATES toolbox [21] and runs on the same platform. One run of 3000 generations takes roughly four hours (also real-time).

Conclusions

In this paper we have described a hybrid expert system for the interpretation of NMR spectra of proteins. Although spectrum interpretation is an iterative process, in which a partial assignment can lead to a more complete assignment, it is important to extract as much information as possible from the spectra automatically, so that the expert can concentrate on the really complicated parts. The results indeed show that the trial-and-error process that constitutes the spectrum interpretation can not be discarded by the use of such systems, but can be significantly enhanced by the combination of heuristics and powerful search methods. Using a global set of settings, 47 patterns were assigned correctly in the three proteins without any human intervention, more than a quarter of all patterns. If for each protein an optimized set of settings is used, the number increases to 62 patterns, more than 35%. This illustrates the power of the approach. In real, life, however, performance will be probably much higher since additional information from the spectra can be used by a human expert to solve ambiguities. A simulation of that situation on the protein BPTI yielded a performance of 77%. Therefore, we believe that these systems will form a great help in the spectrum interpretation.

The above results can be further improved in a variety of ways, each with its own advantages or disadvantages. Information on peak shapes and intensities can be taken into account, NOESY spectra with different mixing times can be used to distinguish between long-range and short-range NOEs, extra rules can be added to assign protons beyond the C_γ-positions and thus facilitate the assignment to amino acid classes, and more runs with the genetic algorithm can be started simultaneously. Furthermore, a feed-back loop that incorporates sequential information in the earlier stages will significantly improve the results of the system. Also, it is important to train the expert system on more proteins than the three that have been used so far, and preferable train the system on experimental spectra of high quality. This way, optimal results can be obtained in the future.

REFERENCES

1 P. Catasti, E. Carrara and C. Nicolini, Comput. Chem., 11 (1990) 805.
2 G.J. Kleywegt, R.M.J.N. Lamerichs, R. Boelens and R. Kaptein, J. Magn. Reson., 85 (1989) 186.
3 F.J.M. van de Ven, J. Magn. Reson., 86 (1990) 633.
4 M. Billeter, V.J. Basus and I.D. Kunz, J. Magn. Reson., 76 (1988) 400.
5 C. Cieslar, G.M. Clore and A.M. Gronenborn, J. Magn. Reson., 80 (1988) 119.
6 P.L. Weber, J.A. Malikayil and L. Müller, J. Magn. Reson., 82 (1989) 419.
7 L. Fu, Connect. Sci., (1989) 325.
8 S.I. Gallant, Comm. ACM, 31 (1988) 152.
9 D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
10 P.G. Politakis, Empirical Analysis of Expert Systems, Pitman, London, 1985.
11 A. Ginsberg, Automatic Refinement of Expert System Knowledge Bases, Pitman, London, 1988.
12 C.B. Lucasius and G. Kateman, in R. Männer and B. Manderick (Eds.), Proc. 2nd Workshop on Parallel Problem Solving from Nature, 28–30 Sept. 1992, Brussels, Elsevier, Amsterdam, 1992, pp. 239–247.
13 K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, New York, 1986.
14 G.M. Crippen, Distance Geometry and Conformational Calculations, Research Studies Press, Chichester, 1981.
15 A.D. Kline and K. Wüthrich, J. Mol. Biol., 192 (1986) 869.
16 G. Wagner and K. Wüthrich, J. Mol. Biol., 155 (1982) 347.
17 F.J.M. van de Ven and C.W. Hilbers J. Mol. Biol., 192 (1986) 389.
18 F.J.M. van de Ven and C.W. Hilbers, J. Mol. Biol., 192 (1986) 419.
19 R. Wehrens, L. Buydens and G. Kateman, Chemom. Intell. Lab. Syst., 12 (1991) 57.
20 K.H. Gross and H.R. Kalbitzer, J. Magn. Reson., 76 (1988) 87.
21 C.B. Lucasius and G. Kateman, GATES: Genetic Algorithm Toolbox for Evolutionary Search, 1991, Software Library in ANSI C, Laboratory for Analytical Chemistry, Catholic University, Nijmegen.

Analytica Chimica Acta, 277 (1993) 325–331
Elsevier Science Publishers B.V., Amsterdam

Information theory of chromatography and titration

Yuzuru Hayashi and Rieko Matsuda

National Institute of Hygienic Sciences, 1–18–1 Kami-yoga, Setagaya, Tokyo 158 (Japan)

(Received 12th October 1992; revised manuscript received 5th February 1993)

Abstract

The function of mutual information, abbreviated as FUMI, is a function of peak parameters (width, area and position, if the peak is Gaussian), separation (resolution *Rs*) and noise variance of an output of an analytical system (e.g. chromatogram). For quantitative analysis, FUMI describes the precision (or relative standard deviation) of measurements as the Shannon mutual information. FUMI theoretically confirms that a high, sharp, well-separated peak gives a reliable measurement (e.g. concentration). The higher the FUMI for a peak, the more precise is the measurement obtained from the peak. This paper demonstrates that chemically different analytical methods intended for quantification, LC and titration, can be evaluated in a general way with FUMI as a criterion.

Keywords: Liquid chromatography; Titrimetry; FUMI; Information theory; Mutual information function

Many analytical systems intended for quantification provide electric signals in peaks. Included among them are chromatography, electrophoresis, flow injection analysis, etc. The role of chemistry in these analyses is to improve peak separation and to change the peak shape and area so that the peaks can provide more useful information. For example, in reversed-phase liquid chromatography, the volume fraction of an organic component in an aqueous mobile phase is changed so that the peaks are more separate and sharper. The selection of the appropriate detection wavelength is useful to increase the peak areas for target materials and to give a higher signal/noise ratio. This process is known as optimization [1–4].

The chemistry to modify the peak shape and separation is very important, because, a sharp, high, well-separated peak leads to accurate and precise measurements (e.g. concentration). However, in practice, it is not always feasible to obtain these statistical quantities by doing experiments, especially when the optimum conditions are searched for among many operating conditions. This is because a large number of experiments should be carried out for the same samples under each set of the operating conditions to obtain statistically valid results. Then, it is fervently desired to discover a mathematical description of the statistical quantities as a function of the peak parameters (area, width, separation, etc.).

According to the theory of the function of mutual information, abbreviated as FUMI Φ, if an analytical system has peak-shaped signals in its output, the precision of the system can be obtained theoretically for each set of operating conditions without the tiresome repetition of experiments [2]. FUMI is a function of peak width, area, separation and noise level and describes the precision (or relative standard deviation) of measurements as the mutual information.

Another purpose of quantitative analysis is rapidity. The efficiency ϑ of analysis has been introduced to express the rapidity numerically as

Correspondence to: Y. Hayashi, National Institute of Hygienic Sciences, 1–18–1 Kami-yoga, Setagaya, Tokyo 158 (Japan).

the average transmission speed (Φ/t_q) of the information FUMI (Φ) during the run time (t_q) [2]. In the LC optimization for antipyretics determination, the most precise analysis (characterized by the maximum of the information, Φ) and the most efficient (rapid) analysis (by the maximum of the efficiency, ϑ) has been demonstrated for the six chromatographic variables (mobile phase composition, wavelength, etc.) [5,6]. The statistical reliability of these optima has been verified experimentally [5].

In chromatography, other sources of error such as injection volume and temperature control are dominant in practice. However, a good injector will always work well for any sample and can be considered independent of the chemical properties of analytes. On the other hand, the chemistry of chromatography (mobile phase composition, etc.) needs to be modified from sample to sample to discover the optimum operating conditions of analysis as mentioned above. Because of this independence, FUMI theory treats only the sample-dependent chemistry which affects peak overlap, shape and noise level. The whole optimization considering the injection error, mobile phase composition, etc. would yield the same results as the FUMI-based optimization. The variables such as injection volume are called evironmental parameters and those concerning the chemistry of chromatography are called controllable variables [2].

The precision is the only statistical parameter that FUMI treats. Accuracy is not included in FUMI, because the theory of Kalman filter does not take it into account. However, the accuracy and precision have been shown to correlate closely with each other and to follow a similar tendency toward peak overlap and peak area ratio in chemical experiments [7] and computer simulations [8,9]. In the optimal experimental condition where FUMI is a maximum, the accuracy will usually also be very high.

Along these lines, we can see that many quantitative analyses can be evaluated generally according to the Shannon information FUMI which refers to the peak shape, separation and noise level in output of the systems. LC and titration are considered in this paper.

THEORY

Precision of data

A brief review of FUMI is given, especially in chromatographic situations. The differentiated titration curves have the same statistical properties as chromatograms: a sharp, smooth, high, well-separated peak indicates high precision. Even one peak can provide analytical information, because we can obtain a measurement such as concentration from a single peak in an output (chromatogram) of an analytical system. First, the Shannon mutual information (FUMI) of a peak without overlap is examined. The effect of peak overlap will be considered afterward.

Let us assume that the peak is reproducible under the same operating conditions except the white noise. This situation mimics the repeated experiments on the same samples under the same operating conditions. We can obtain a measurement from each experiment by using least-squares curve fitting such as the Kalman filter. Finally, we can calculate the statistical quantities, average and standard deviation by repeating the same experiments.

According to the theory of Kalman filter, the relative standard deviation (R.S.D.$_j$) of these measurements can be described by the peak shape $F_i(j)$ and noise level $\tilde{W}_c$ under the above assumption [2]:

$$1/\text{R.S.D.}_j = \frac{\sum_{i=\alpha}^{\beta} F_i(j)^2}{\tilde{W}_c} \tag{1}$$

where $F_i(j)$ denotes the signal intensity of peak j at data point i. $\tilde{W}_c$ denotes the variance of the white noise in the output.

If the signal is a Gaussian peak with area A_j and width σ_j, we can obtain the following equation by simple mathematics from the above equation [2]:

$$1/\text{R.S.D.}_j = \left[\frac{A_j^2}{2\pi^{1/2}\sigma_j\tilde{a}}\right]^{1/2} \tag{2}$$

where $\tilde{a}$ denotes the product of the sampling interval (and integration period) of an A/D converter (ΔT) and the noise variance (W_c): $\tilde{a} = \Delta T W_c$

Let the precision be the reciprocal of the relative standard deviation. According to information theory, the information for peak j can be described as the logarithm of the precision [2]:

$$\Psi_j = \log(1/\text{R.S.D.}_j) \quad (3)$$

The most prominent advantage of this equation is that the statistics (precision) of a method can be calculated from the peak area, width and noise level in an output of an analytical system without repeated experiments or computer simulations (for peak overlap, see Eqn. 4).

If peak j overlaps with any other peak, the information Ψ_j is decreased by the information loss $\delta\phi_j$ [2]:

$$\phi_j = \Psi_j - \delta\phi_j \quad (4)$$

ϕ_j is the information that we can actually collect from the real signals for peak j and is called the mutual information. The information loss $\delta\phi_j$ can be derived by the Tailor series expansion of the information ϕ_j and is a function of the positions, areas and widths of peak j itself and adjacent peaks $j-1$ and $j+1$. Roughly, the loss $\delta\phi_j$ is a function of the resolution Rs.

From information theory, the information of signals independent of each other is additive. Then, the total information Φ of multi-peaks in an output of a system is equal to the sum of the individual bit of information ϕ_j [2]:

$$\Phi = \sum_{j=1}^{q} \phi_j \quad (5)$$

where q denotes the number of analytes. Equations 4 and 5 are called FUMI.

The intensity of the white noise varies from data point to data point and from experiment to experiment, but the above equations only include the quantities which are assumed to be constant in the situations of the repeated experiments. That is, given the peak shape, position and noise variance, we can theoretically predict the standard deviation, precision and information. In practice, we can calculate these statistical quantities by estimating the peak shape and noise variance from a single chromatogram. We do not have to repeat the experiments on the same samples under the same operating conditions to calculate these statistical quantities, if the experiments are considered reproducible.

To summarize, the most important feature of the information FUMI is that it refers to individual peak information and is a function of peak shape (area and width), overlap and noise level. For example, if the peak area is kept constant, the information ϕ_j increases with decreasing the width σ_j or peak sharpening (see eqn. 3). FUMI is more reduced by the peak overlap than by changes in the peak shape (broadening). This is the reason why the peak resolution is extremely important to chromatographic analyses.

Efficiency of analysis

A numerical description of the rapidity of analysis is the analytical efficiency ϑ which is defined as the average transmission speed (Φ/t_q) of the information Φ provided by an analytical system. t_q denotes the analysis time. As long as the information Φ takes a large value, a high efficiency ϑ means rapid analysis with satisfactory precision.

The most efficient analysis which is characterized by the maximum of the efficiency ϑ coincides with the minimum-time analysis in liquid chromatography in many cases. However, even if the observation time for the minimum-time analysis is less than the observation time for the most efficient analysis, more information can be obtained from the latter case in a fixed time period. This is the reason why the efficiency ϑ is preferred here rather than the observation time t_q.

The consumption of organic solvent in a mobile phase of liquid chromatography poses problems associated with environmental pollution. If the efficiency is defined as (total information)/(total volume of solvent used), the conditions of the lowest solvent use will be selected by the FUMI-based optimization. The cost–performance characteristics of chromatography can be evaluated similarly. We can safely say that the analytical efficiency ϑ is a general description of the rapidity, conservation and economy of the analysis.

OPTIMIZATION OF LC ANALYSIS FOR DRUGS

With FUMI as a criterion, the following six variables were optimized for the determination of a mixture of antipyretics over the practical working ranges [5]: acetonitrile volume fraction in aqueous mobile phase X, column length L, mobile phase velocity u, detection wavelength λ, the selection of the best internal standard I_s and its amount $m_s(I_s)$. The detailed information about the X dependence of the capacity factor of each analyte drug, ultraviolet spectrum, etc. is given elsewhere [5]. Now, the precision Φ and efficiency ϑ are the functions of the six variables to be optimized. This simultaneous optimization of many variables is called TOCO (total chromatographic optimization).

Figure 1 shows the chromatograms of the precision optimum (A) and the efficiency optimum (B) [5]. The selected values of the variables (optima) are described in Table 1. The precision optimum gives the maximum of FUMI among all the examined conditions and has the smallest theoretical R.S.D. (0.02%; derived from the FUMI value (maximum) in Table 1) but a long observation time (ca. 8 min). The efficiency optimum is characterized by the maximum of the efficiency ϑ and has the shortest run time (ca. 50 s) and satisfactory precision (theoretical R.S.D. = 0.06%).

TABLE 1

Conditions [a] for the precision optimum and efficiency optimum

Variable	Optima	
	Precision	Efficiency
L (cm)	5	5
u (mm s^{-1})	0.5	5
X (%)	18	194
λ (cm)	280	280
I_s	6	6
m_s (μg ml^{-1})	9.6	9.6
Φ	34.37	29.81
ϑ	0.0765	0.716
Analyte [b]	Observed R.S.D. (%) (n = 8)	
1	0.12	0.26
2	0.24	0.33
3	0.23	0.35

[a] Detailed experimental conditions are given elsewhere [5].
[b] Analytes: 1 = acetaminophen; 2 = caffeine; 3 = salicylamide.

The observed errors were excellent for the TOCO [5]: 0.12–0.24% for the precision optimum and 0.26–0.35% for the efficiency optimum (see Table 1). The errors are underestimated due to the fact that FUMI neglects the error sources

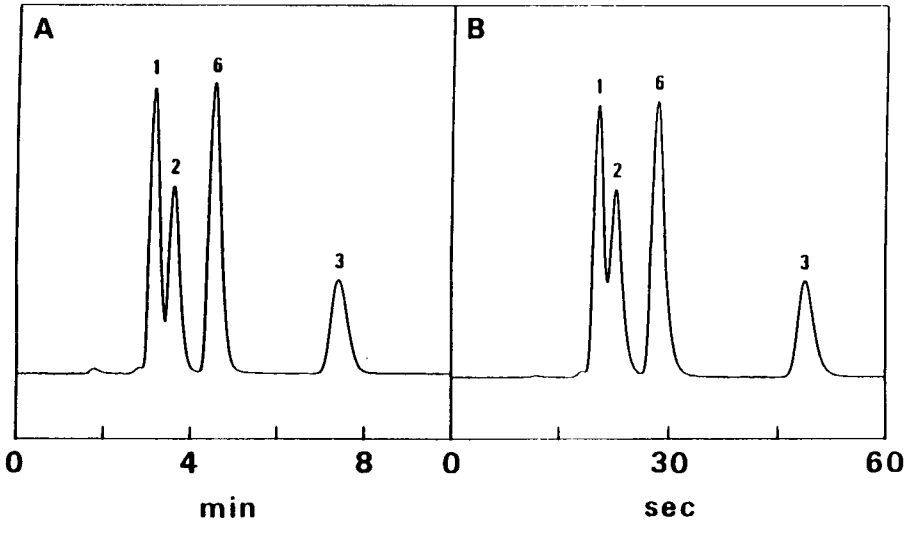


Fig. 1. The precision optimum chromatogram (A) and the efficiency optimum chromatogram (B) ($\hat{Rs}$ = 1.03). Peak 1 denotes acetaminophen; peak 2, caffeine; peak 3, salicylamide; peak 6, 8-chlorotheophylline (the best internal standard selected in the TOCO). The detailed conditions selected are listed in Table 1. (From Ref. 2, by permission of Elsevier.)

TABLE 2

Comparison between the theoretical and observed R.S.D. values [a]

Sample (μg ml^{-1})	R.S.D. (theoretical) (%)	R.S.D. (observed) (%)
1 (ethyl paraben)	0.073	1.81
1 (propyl paraben)	0.12	1.80
1 (butyl paraben)	0.11	4.13
0.02 (ethyl paraben)	3.65	43.5
0.02 (propyl paraben)	6.15	42.2
0.02 (butyl paraben)	5.71	53.0

[a] The R.S.D. values were obtained by repeated experiments ($n = 10$) on the same samples under the same operating conditions: column, Develosil ODS-K3 (100 mm×4.6 mm i.d.); eluent, acetonitrile–water (4:6); flow-rate, 1 ml min^{-1}; detection, UV 250 nm; injection volume, 5 μl. Detailed experimental conditions will be published.

which may prdominate in practice as mentioned in the introductory section.

FUMI selects as the precision optimum the sharpest possible peaks (the highest S/N ratio) among the chromatograms which show good separation. The efficiency optimum is the most rapid analysis among the conditons which show good separation and a good S/N ratio.

Table 2 shows the correlation between theoretical and observed RSDs for HPLC determination of ethyl paraben, propyl paraben and butyl paraben. As the amount of sample to be quantified is decreased, the variance of the white noise increases proportionately and predominates over the other error sources. In this situation, the estimated errors based on FUMI approach the observed ones (compare the R.S.D.s for 1 and 0.02 μg ml^{-1}). That is, in noisy chromatograms often appearing in trace analyses, FUMI gives a better estimation of all the chromatographic errors.

FUMI has an arbitrary constant $\hat{R}s$ of the resolution $\hat{R}s$. It is assumed that the data processing used in the system cannot provide exact results below the limit $\hat{R}s$. In Fig. 1, $\hat{R}s$ is 1.03. Peaks 1 and 2 overlap slightly in the efficiency optimum chromatogram. This optimum holds for data processing which has high peak-resolving power such as the Kalman filter. If a large $\hat{R}s$ (e.g. 1.53) is input in FUMI, another efficiency optimum of more separated peaks is recommended by the TOCO (see Fig. 2) [2]. The perpendicular dropping is considered to be of inferior power [10], but can provide reliable results from this optimal chromatogram. In general, the lower limit $\hat{R}s$ of the resolution depends on the areas of the adjacent peaks [11] and is set here for the peaks of the same areas.

The analytical role of a variable may be regarded as how the variable affects the ultimate purposes of the analysis, the precision and efficiency (rapidity) [2,6]. This analytical role can be recognized by a plot against the precision Φ and efficiency ϑ called the Φ–ϑ plot. A chromatogram or a set of operating conditions is identified as a point in the space spanned by the precision Φ and efficiency ϑ, because it has its own numerical values of Φ and ϑ. Then, the analytical role of a variable can be represented by its own curved line in the space when it is changed during the one-variable optimization. In Fig. 3, the solid lines denote the experimental conditions or chromatograms which show good separation ($\hat{R}s > 1.53$ for each peak pair) and the dotted lines denote poor separation ($\hat{R}s < 1.53$).

From the slopes of the Φ–ϑ plots for the mobile phase composition and column length (see

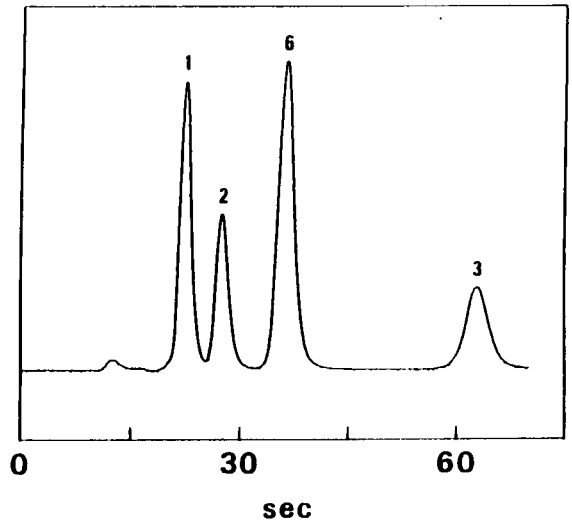


Fig. 2. The efficiency optimum chromatogram with another $\hat{R}s$ (1.53). (From Ref. 2, by permission of Elsevier.)

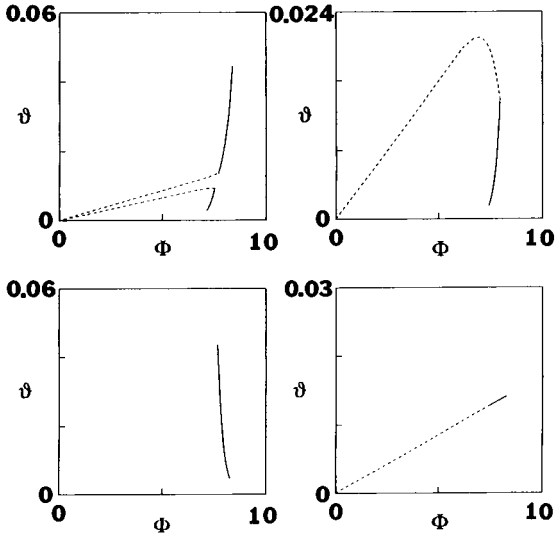


Fig. 3. The Φ–ϑ plots of (A) mobile phase composition, (B) column length, (C) mobile phase velocity and (D) plate number. These Φ–ϑ lines were calculated on the basis of Figs. 3 and 5 of Ref. 2.

the solid lines of Fig. 3A and 3B), we can obtain at once the precision and rapidity by modifying these variables, if the peaks are separated far more than necessary. The dotted lines represent the bad conditions where the log(capacity factor) vs. X lines are crossing (Fig. 3A) or the column is too short to resolve the peaks sufficiently (Fig. 3B). On the other hand, the Φ–ϑ plot for the mobile phase velocity provides an example of how precision or efficiency can be obtained at the expense of the other (see the negative slope of the Φ–ϑ line of Fig. 3C).

The plate number of a column N is often called the column efficiency. Figure 3D shows that the analytical efficiency ϑ increases as the column efficiency increases. An increase in N means an improvement in the stationary phase without changing any other variable such as column length, etc. This Φ–ϑ plot is a straight line, because the plate number does not affect the analysis time. The column length also changes the plate number. The different analytical roles of the plate number and column length can be recognized from the Φ–ϑ lines of Figs. 3B and 3D. However, this difference cannot be distinguished by the resolution $\hat{R}s$ [2,6].

The above optimization is based on a complete survey of every chromatogram over the examined ranges of the variables. However, FUMI is applicable to the simplex method [12,13] which dispenses with the evaluation of all the response surfaces. This is because FUMI greatly increases if a new peak is discovered from a peak which appeared to be a single large peak from the previous experiment [14].

EVALUATION OF TITRATION

The efficiency of titration can also be calculated, but it will not be so important as the

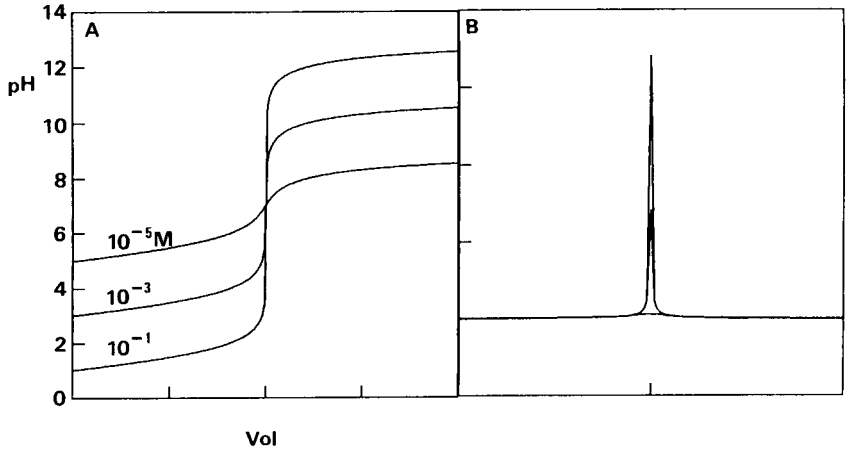


Fig. 4. The titration curves of (A) strong acid with strong base, and (B) the differentiated curves.

TABLE 3

The theoretical and observed errors [a] for titration

Concentration (M)	Φ	R.S.D. (%)	R.S.D.1 (%)	R.S.D.2 (%)
0.1	5.83	0.29	0.00001	0.0044
0.01	5.49	0.41	0.00006	0.0062
0.001	4.98	0.69	0.0011	0.013
0.0001	4.06	1.72	0.0060	0.050
0.00001	2.71	6.67	0.012	3.30

[a] R.S.D.1 and R.S.D.2 (relative standard devation) were obtained from the repeated simulations ($n = 10$). The theoretical R.S.D. and information Φ were calculated from the derivative curves.

precision of this analysis, which is in sharp contrast to the chromatographic situations.

Figure 4 shows the typical example of the acid–base titration. The curves of Fig. 4A were obtained from the simulation of a titration of a strong acid (e.g. HCl) with a strong base (e.g. NaOH). Figure 4B shows the first-derivatives of the titration curves. As the concentration of the acid decreases, the degree of the pH jump around the equivalent point decreases and the peak area of the first-derivative curves decreases. At the concentration of 10^{-5} the peak of the derivative curve is so small that it is difficult to recognize it by visual inspection.

A high concentration of the acid is known to give an exact data. This fact is well explained by FUMI: the first-derivative curve of a high concentration gives a large amount of information FUMI or high precision. Table 3 lists the relative standard deviation (R.S.D.) of the measurements obtained from the simulated noisy data. Column R.S.D.1 shows the R.S.D. values obtained by the least-square curve fitting (damping Gauss–Newton) of the titration curves and the column R.S.D.2 the R.S.D. values by determining the zero point from the second derivative of the curves. The lower R.S.D. for the R.S.D.1 comes from a larger number of data points used in the data processing than R.S.D.2. As the information FUMI decreases, the precision decreases (i.e. the R.S.D. increases).

The titration of a weak acid of various pK_as and the titration of the mixture of two acids can be simulated similarly. The improvement in the titration methods (e.g., reverse titration, etc.) can be interpreted in terms of FUMI.

Conclusion

The optimization of LC analysis and the evaluation of titration have been presented. It is extremely interesting that the chemically different methods, titration and chromatography, can be evaluated or optimized by the single criterion FUMI. We believe that the statistical reliability of many analytical methods such as chromatography, electrophoresis, titration, FIA and gravimetry can generally be evaluated by FUMI.

REFERENCES

1 D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: a Textbook, Elsevier, Amsterdam, 1988.
2 Y. Hayashi and R. Matsuda, Chemom. Intell. Lab. Syst. 18 (1993) 1.
3 P.J. Schoenmakers, Optimization of Chromatographic Selectivity, Elsevier, Amsterdam, 1986.
4 K. Jinno, Computer-Assisted Chromatography System, Hüthig, Heidelberg, 1990.
5 Y. Hayashi, R. Matsuda and A. Nakamura, Chromatographia, 30 (1990) 85.
6 Y. Hayashi and R. Matsuda, Chromatographia, 30 (1990) 171.
7 Y. Hayashi, T. Shibasaki, R. Matsuda and M. Uchiyama, Anal. Chim. Acta, 222 (1987) 187.
8 Y. Hayashi and S.C. Rutan, Anal. Chim. Acta, 271 (1993) 91.
9 Y. Hayashi, R.S. Helburn and S.C. Rutan, Proc. 4th Symp. Comput.-Enhanced Anal. Spectrosc., 1992, in press.
10 Y. Hayashi, T. Shibasaki, R. Matsuda and M. Uchiyama, J. Chromatogr., 407 (1987) 59.
11 R. Mautsuda, Y. Hayashi, T. Suzuki and Y. Saito, Chromatographia, 32 (1991) 233.
12 J.C. Berridge, Chemom. Intell. Lab. Syst., 3 (1988) 175.
13 J.C. Berridge, Chemom. Intell. Lab. Syst., 5 (1989) 195.
14 R. Matsuda, Y. Hayashi, T. Suzuki and Y. Saito, J. Chromatogr., 585 (1991) 187.

Analytica Chimica Acta, 277 (1993) 333–346
Elsevier Science Publishers B.V., Amsterdam

Taxonomy of *Amanita* mushrooms by pattern recognition of amino acid chromatographic data

Robert L. White, Peter D. Wentzell and Mark A. Beasy

Department of Chemistry, Dalhousie University, Halifax, Nova Scotia B3H 4J3 (Canada)

Debra S. Clark

Department of Chemistry, Acadia University, Wolfville, Nova Scotia B0P 1X0 (Canada)

Darryl W. Grund

Department of Biology, Acadia University, Wolfville, Nova Scotia B0P 1X0 (Canada)

(Received 3rd September 1992; revised manuscript received 19th November 1992)

Abstract

Pattern recognition methods have been applied to the classification of six species of fungi of the genus *Amanita* based on their amino acid content. The six species investigated were *A. flavoconia*, *A. flavorubescens*, *A. gemmata*, *A. muscaria*, *A. rubescens*, and *A. virosa*. A total of 59 samples were examined and 66 unique peaks were found. Liquid chromatographic methods are described and the mean concentrations of 15 identified amino acids in each species are given. Principal components analysis and hierarchical clustering were applied to the classification problem using six identified amino acids. Peak areas normalized for the total amino acid content were found to be the best descriptors. Complete separation of four species was possible with these methods. The remaining two species (*A. muscaria* and *A. rubescens*) formed relatively tight clusters in principal components space, but were overlapped with other species.

Keywords: Chromatography; Pattern recognition; *Amanita* mushrooms; Amino acids; Fungi; Mushrooms

Pattern recognition methods have become well-established tools for the analysis of multivariate chemical data sets. Over the years, these techniques have been applied to problems of classification in a wide variety of fields [1,2] and the areas of application continue to grow as the methodologies become more broadly recognized and readily available. The utility of these methods arises not only from their ability to classify unknown samples, but also from their capacity to elucidate how various chemical features contribute to class distinctions, thereby leading to a better understanding of fundamental differences among classes. One area where the application of multivariate methods has been particularly successful is in the classification of biological species through chemical markers, or chemotaxonomy [3–6]. As an alternative to classification based on morphological or biometric features, chemotaxonomy may be a more reliable or efficient tool in certain cases. Differences in the chemical makeup of biological organisms is a natural consequence of molecular evolution. Chemotaxonomic approaches may examine similarities between genetic sequences in different species directly, or

Correspondence to: P.D. Wentzell, Department of Chemistry, Dalhousie University, Halifax, Nova Scotia B3H 4J3 (Canada).

focus on the amino acid sequences in selected proteins or enzymes that are coded from the DNA. Alternatively, changes in the enzymatic environment of the organism may manifest themselves through changes in the distributions of other chemical families (e.g., proteins, free amino acids, cuticular hydrocarbons). These differences can be more difficult to interpret since they are less direct, arising from numerous complex biochemical interactions and environmental factors.

In this work, pattern recognition methods are applied to the problem of classification for several species of mushrooms in the genus *Amanita*, a common group of fungi containing both edible and toxic members [7,8]. The free amino acid content of the mushroom samples as determined by liquid chromatography (LC) forms the basis of the study. This problem was of interest for several reasons. First, the distinction of mushrooms by purely morphological means is often a difficult problem and, given the toxicity of certain species, one of practical importance. The variations in the number and concentrations of amino acids from species to species may contribute to the overall toxicity [9]. In some cases, toxicity may be attributed to a particular amino acid, but its identification and detection can pose problems. Differences in the amino acid content of different species might be expected on the basis of the biochemical pathways leading to their synthesis. Product distributions will depend on the abundance of particular enzymes as well as on other internal and external chemical factors. The use of amino acid content for taxonomic purposes in a variety of organisms has been investigated [5]. In 1960, Close [10] reported that amino acid content of eight different fungi could not be used as a basis for their classification. Hatanaka and Terakawa [11] reached a similar conclusion based on the study of 73 fungi, but both of these studies utilized only qualitative or semi-quantitative data from paper chromatography. Catalfomo and Tyler [12] reported qualitative data for the free amino acid content of species of *Amanita* and *Vaginata* and pointed out that a more quantitative study was needed to properly evaluate the taxonomic utility of these compounds. Robbers et al. [13] later found that the distributions of amino acids and other compounds in 39 species of the genus *Inocybe* were useful for classification, and Tyler et al. [14] demonstrated that semiquantitative measurements on the urea content of 344 species of fungi had some taxonomic value. Cohen and Farnham [15] provided quantitative data on the amino acid content of the fungus *Phycomyces blakesleeanus* and expressed doubts about the taxanomic utility of such data due to the dependence of free amino acid pools on environmental factors. Faulstich et al. [16] have suggested using virotoxins to distinguish *A. virosa*, which is deadly poisonous, from the morphologically very similar *A. bisporigera*. The first objective of the work presented here was to determine if a more rigorous analysis of quantitative data for the amino acid content of mushrooms would reveal differences among species and allow classifications to be made. A second objective was to determine if such classifications revealed noteworthy relationships among the species or clues regarding differences in the biochemical pathways present in each case. Finally, information on the analytical procedures employed and the mean amino acid content of each species is provided.

EXPERIMENTAL

Reagents

Methanol and tetrahydofuran (BDH, Toronto, ON) were glass-distilled, chromatography grade reagents. Deionized water was obtained from a Barnstead Nanopure cartridge system (Fisher Scientific, Dartmouth, NS) and stored in glass. The derivatization reagent was FLUO-R, System 6300 high-performance fluorescence reagent (Beckman Instruments, Mississauga, ON). The amino acid standards were obtained from Sigma (St. Louis, MO), except L-glutamine and L-lysine (Eastman Kodak, Rochester, NY), L-tryptophan and L-glutamic acid (Fisher Scientific), DL-aspartic acid (Aldrich, Milwaukee, WI), L-tyrosine (ICN, Cleveland, OH), and DL-valine (Nutritional Biochemicals, Cleveland, OH); all were of analytical grade. Sodium acetate was reagent grade (Anachemia, Montreal, QC).

Fungal samples

The fungal samples of *Amanita* were collected during September to November 1985, and July to October 1986, primarily in Kings, Hants and Annapolis Counties, Nova Scotia. The data set which forms the basis of this study was generated from 59 samples of mushrooms from six species: *A. flavoconia* (poisonous), *A. flavorubescens* (poisonous), *A. gemmata* (poisonous), *A. muscaria* (poisonous/hallucinogenic), *A. rubescens* (nonpoisonous) and *A. virosa* (deadly poisonous). Ten samples of each were used except for *A. flavorubescens*, for which only 9 were available. The geographic distribution of samples is indicated in the map in Fig. 1. The number of samples of each species collected in each locality is given next to the corresponding symbol. In cases where multiple samples were collected from one area, different sites were often used. As much as possible, the collection procedure attempted to remove correlations among dates and locations with species, but to some extent these are restricted by preferred habitats and growing seasons. Photographs, spore prints, and macroscopic descriptions were taken for representative specimens. Spore prints were obtained by placing the pileus (cap) on white paper for several hours and were stored in plastic wrap at room temperature. All collections were air dried on the same day they were obtained by placing them on wire screens over an electric fan-heater at 35°C for 48 h. Following this, they were rewrapped in waxed paper and stored at room temperature in a ventilated cabinet in proximity to a pest control strip to prevent damage by insects.

Microscopic examinations of representative fruit bodies were performed on fresh or rehydrated (reconstituted) dried samples of each *Amanita* species. Rehydration was accomplished by immersing a pileal wedge in 75% (v/v) ethanol, blotting it dry, then placing it in distilled water. Thin sections of pileus and lamellae were then cut from the rehydrated wedge, stained with 1% (m/v) aqueous phloxine B ($C_{20}H_2Br_4Cl_4Na_2O_5$, Baker, Phillipsburg, NJ) for 1.5–2.0 min, blotted, mounted in 3% (m/v) potassium hydroxide, and viewed with an optical light microscope. Details of pileal cuticle type, pileal context and lamellar tramal type were recorded. From the spore prints, mounts were made in Melzer's reagent (1.5 g KI plus 0.5 g I_2 dissolved in 20 ml distilled water plus 20 ml 1:1 (m/v) chloral hydrate in distilled water) and details of spore size, shape and amyloidity were recorded.

Analytical methods

Prior to chromatographic analysis, each sample of dried mushroom cap (between 0.090 and 0.850 g) was subjected to a one minute extraction into 75 ml of 80% (v/v) ethanol–water with a Polytron PT10-35 ultrasonic homogenizer (Brinkman Instruments, Rexdale, ON), filtered, and diluted to 100 ml with 80% ethanol. The mushroom stems were not used in the analysis to ensure greater sample homogeneity. Before injection onto the LC column, 40 μl of the commercial *o*-phthalaldehyde (OPA) derivatizing reagent was added to a mixture of the above filtrate (20 μl) and one of the following: (a) 0.1 M sodium acetate (20 μl, pH 7.2), (b) a mixture of internal standard solution (10 μl, homoserine, 106 μM) and 0.1 M sodium acetate (10 μl), or (c) a mixture of internal standard (10 μl) and a spiking solution of standard amino acids (10 μl, ranging from 17–78 μM). Combination (a) was used for blanks, combination (b) for most samples, and combination (c) for peak matching purposes. After 1 min at room temperature, the derivatization reaction was quenched by the addition of 0.1 M sodium acetate (120 μl) and 20 μl of this solution was injected onto the chromatograph.

The chromatographic separation of the fluorescent isoindole derivatives was based on methods previously pubished [17,18]. The procedure employed a Beckman high-performance liquid chromatograph (Beckman Instruments) fitted with reversed-phase column (5 μm Spherisorb ODS2 C_{18}, fully end-capped, 150 × 4.6 mm i.d. Chromatography Sciences, Montreal, QC) and an Upchurch C-130B guard column (20 × 2 mm) packed manually with Perisorb RP18 (30–40 μm, Chromatography Sciences). Gradients were generated by the proportional pumping of two solvents: 0.1 M aqueous sodium acetate–methanol–tetrahydrofuran (900:95:5, v/v/v) and methanol. Solvents were degassed and the total flow-rate was

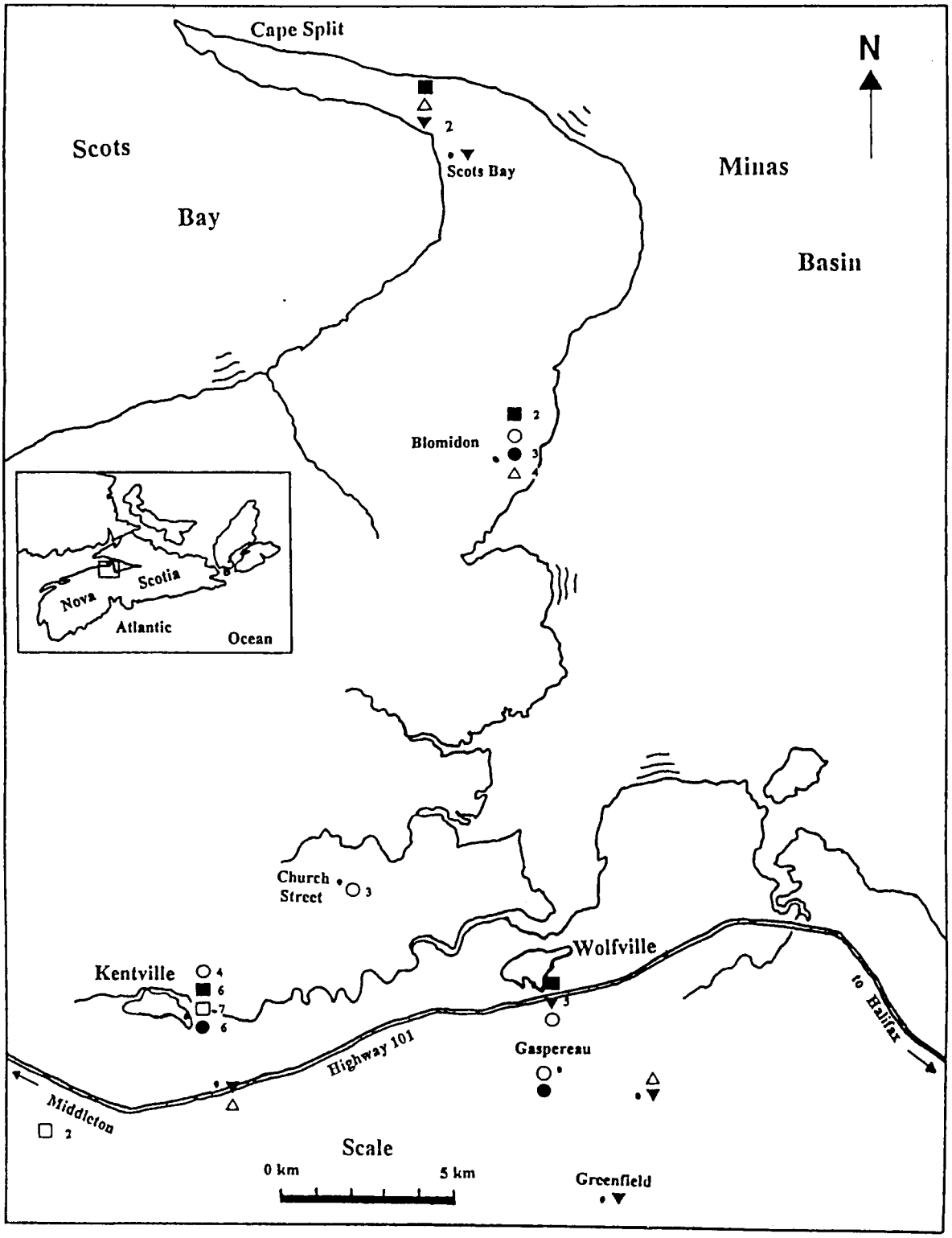


Fig. 1. Geographic distribution of samples. A map of the province of Nova Scotia is inset showing the location of the region. Legend: □ = *A. flavorubescens*; ● = *A. virosa*; ■ = *A. flavoconia*; ○ = *A. rubescens*; ▼ = *A. gemmata*; △ = *A. muscaria*. Numbers refer to number of samples (no number means one sample).

maintained at 1.2 ml min^{-1}. The solvent program was optimized to give satisfactory separation of the derivatives of interest within 35 min and is given below:

Time (*min*)	0.00	1.04	12.08	17.08	20.00
Methanol (%)	0	16	16	32	32

Time (*min*)	22.08	27.08	30.00	32.08	35.00
Methanol (%)	60	60	100	100	0

A typical chromatogram also showing the gradient composition is presented in Fig. 2.

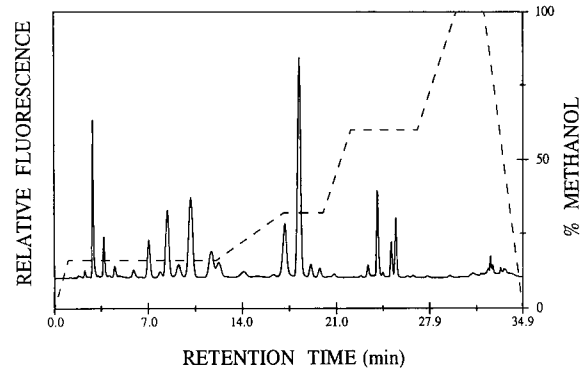


Fig. 2. Typical chromatogram of derivatized mushroom extract (*A. virosa*) showing gradient program.

Dilutions of a stock solution containing 16 standard amino acids (L-alanine, 1.17 mM; L-arginine, 0.60; L-asparagine, 0.60; DL-aspartic acid, 0.49; L-glutamic acid, 0.35; L-glutamine, 0.77; glycine, 1.41; L-histidine, 0.99; L-isoleucine, 1.20; L-leucine, 1.30; L-lysine, 1.55; L-phenylalanine, 0.93; L-serine, 0.86; L-threonine, 0.86; L-tryptophan, 0.49; L-tyrosine, 0.57) were used to determine the reproducibility of the method (ten injections of the stock diluted 1:20). To prepare peak area calibration curves, six concentrations over the range of 7.5–200 μM were used. The homoserine internal standard solution, injected with the amino acid standard, was diluted to match the concentration of the other amino acids. Reproducibility and calibration data are given in Table 1.

The major peaks in one extract per species were identified by co-injection of standard amino acids with the extract. Four separate amino acid standard mixtures were employed for spiking: (a) L-isoleucine, L-serine, L-threonine, L-tryptophan; (b) L-alanine, L-arginine, DL-aspartic acid, L-glutamine, L-methionine, L-leucine; (c) L-glutamic acid, L-histidine, L-ornithine, L-tyrosine, DL-valine;

TABLE 1

Retention, reproducibility and calibration data for the standard amino acid mixture

Amino acid	Retention time (min)		Corrected peak area [a]		Concentration (μM)	Relative sensitivity	R^2
	Mean	R.S.D. (%)	Mean	R.S.D. (%)			
Asp	3.15	2.3	29.5	3.8	48.8	0.99	0.99
Glu	4.05	2.5	20.9	3.9	34.7	1.00	1.00
Asn	6.67	3.0	26.5	3.6	59.8	0.85	1.00
Ser	8.10	3.1	39.6	3.4	85.6	0.92	0.99
Gln	9.74	3.3	32.0	3.9	77.3	0.95	1.00
His	10.70	3.5	28.5	5.3	98.6	0.58	0.99
Hse	11.98	3.5					
Gly	13.86	3.6	61.5	3.1	141.2	0.85	0.99
Thr	14.65	3.0	39.6	3.1	85.6	0.96	0.99
Arg	17.72	1.9	34.6	5.3	60.3	0.89	0.96
Ala	19.46	1.4	51.2	3.1	116.7	0.91	0.99
Tyr	20.66	1.4	26.8	3.3	56.9	0.84	0.97
Trp	24.30	0.4	20.5	3.7	49.0	0.85	0.98
Phe	24.91	0.4	38.3	4.2	92.6	0.85	0.99
Ile	25.70	0.4	56.0	3.5	120.5	0.93	0.99
Leu	26.06	0.5	52.7	6.1	129.6	0.92	0.99

[a] Corrected for the internal standard, homoserine (Hse).

TABLE 2

Summary of peak area statistics (mean corrected peak area, percent relative standard deviation, and number of times observed) by species [a]
(There were 10 samples for each species except *A. flavorubescens*, which had 9)

Peak	Amino acid	*A. Flavorubescens* Species 1			*A. Virosa* Species 2			*A. Flavoconia* Species 3			*A. Rubescens* Species 4			*A. gemmata* Species 5			*A. muscaria* Species 6		
		Mean	R.S.D. (%)	No.	Mean	R.S.D. (%)	No.	Mean	R.S.D. (%)	No.	Mean	R.S.D. (%)	No.	Mean	R.S.D. (%)	No.	Mean	R.S.D. (%)	No.
1		6.09	283	1															
2		0.41	96	5	1.66	61	10	1.27	41	10	0.25	153	3	0.99	62	10	2.00	40	10
3					4.83	34	10	0.03	300	1				3.22	72	10	2.07	67	10
4		2.37	87	9	0.17	84	7	6.61	57	10	2.74	82	8	0.77	57	9	1.00	56	10
5								1.64	67	9				0.13	152	4			
6	Asp	55.04	43	9	14.94	40	10	29.18	34	10	53.94	53	10	66.42	39	10	74.40	42	10
7		9.32	283	1	1.60	99	6												
8					0.62	144	9	0.22	62	10									
9					0.01	300	1	0.02	170	3									
10	Glu	84.21	27	9	26.53	33	10	93.34	29	10	96.31	36	10	99.95	30	10	178.28	44	10
11		0.08	283	1	0.69	172	3												
12		1.07	38	9	1.49	75	8	2.69	76	10	2.21	94	9	1.28	89	10	2.36	106	10
13		0.15	123	4				0.98	85	9	1.73	127	5						
14		6.30	62	9	8.46	44	10	5.35	26	10	8.42	59	10	5.43	48	10	12.58	43	10
15		0.42	95	5	0.04	300	1	1.37	127	4									
16														0.10	300	1			
17		2.14	56	8	1.41	105	10	8.05	52	10	5.77	101	8	7.83	39	10	14.71	38	10
18		9.51	283	1															
19																	0.04	300	1
20	Asn	72.05	32	9	13.50	56	10	58.90	43	10	68.68	52	10	45.74	21	10	65.57	40	10
21		10.89	280	2	0.60	100	10	0.98	63	9				0.58	54	8	1.27	30	10
22	Ser	108.73	21	9	49.26	42	10	57.18	44	10	80.84	27	10	92.91	26	10	111.32	27	10
23		35.08	283	1	10.73	86	10	0.54	110	7									
24	Gln	281.41	27	9	55.68	61	10	179.19	51	10	265.73	60	10	260.45	31	10	198.11	42	10
25		8.40	252	2				0.90	200	2	5.31	74	10						
26	His	75.36	44	9	23.64	40	10	96.30	19	10	94.92	38	10	44.18	29	10	52.25	31	10
27		30.75	283	1															
28	Hse																		
29		8.56	283	1															
30	Gly	74.13	18	9	42.77	31	10	31.04	22	10	143.36	201	10	44.54	29	10	57.21	21	10
31	Thr	46.12	26	9	39.13	63	10	54.99	38	10	46.52	40	10	42.70	40	10	52.39	28	10
32		0.62	134	4															
33		0.46	283	1	12.11	35	10	19.43	50	10	5.33	129	5						

34		7.48	31	9	0.70	85	10	27.19	29	10	9.89	94	10	23.51	30	10	20.12	40	10
35								14.24	27	10	2.71	78	9				0.51	162	3
36		15.72	48	9	4.62	68	8	7.82	71	8	24.08	36	10	7.72	47	10	9.63	54	10
37	Arg	133.81	59	9	44.31	30	10	66.62	105	10	86.45	22	10	168.49	66	10	151.10	24	10
38														12.82	113	5	2.25	173	3
39	Ala	338.20	44	9	158.15	32	10	153.14	31	10	283.42	25	10	167.99	32	10	185.58	22	10
40					0.66	153	3												
41		13.22	22	9	15.61	45	10				24.75	56	10	13.09	47	10	19.72	110	10
42	Tyr	15.13	52	9	9.13	32	10	29.53	21	10	12.23	62	10	20.07	55	10	12.99	36	10
43		1.52	153	3															
44														6.97	37	10	4.54	48	10
45					0.21	132	4												
46		21.30	92	9	2.55	69	10	9.69	46	10	6.68	83	10	5.13	43	10	9.68	55	10
47								0.42	294	2									
48					0.17	125	4	0.09	128	4	0.25	300	1						
49		3.77	89	9	1.99	40	10	0.13	283	2	3.06	63	10	2.89	39	10	2.12	36	10
50	EA	7.21	79	9	7.59	44	10	4.53	57	10	8.91	22	10	1.90	68	10	1.55	34	10
51		0.84	100	6	0.11	110	5							0.68	32	10	1.49	49	10
52		0.33	42	9	0.26	30	10	3.18	84	10	3.10	182	6	1.13	64	10	0.64	47	10
53	Trp/Met	17.70	38	9	6.60	47	10	12.21	32	10	9.63	36	10	33.78	39	10	16.18	28	10
54	Met							0.70	137	5									
55	Val	54.06	31	9	17.16	56	10	22.84	38	10	52.99	37	10	28.37	24	10	40.30	38	10
56											0.59	300	1						
57	Phe	16.78	39	9	4.12	48	10	25.55	43	10	14.99	60	10	41.59	32	10	21.04	40	10
58					0.13	123	6	0.62	147	5									
59	Ile	48.17	37	9	23.27	34	10	24.85	21	10	50.99	24	10	27.40	23	10	44.85	27	10
60	Leu	25.87	40	9	4.90	39	10	20.98	46	10	27.99	42	10	29.55	23	10	31.18	32	10
61		0.75	233	3															
62		3.50	63	9	0.98	41	10	0.87	69	10	2.64	56	10	0.95	49	10	2.13	31	10
63	Orn	6.10	49	9	3.09	72	10	9.92	27	10	10.19	63	1	2.85	38	10	4.15	22	10
64		2.87	113	8															
65		1.81	41	9							0.67	300	10						
66	Lys	5.97	35	9	3.75	118	10	12.23	51	10	7.36	59	10	7.52	56	10	8.67	56	10
67		3.05	87	9							5.17	300	1						

[a] Abbreviations: Ala = alanine, Asn = asparagine, Asp = aspartic acid, Arg = arginine, EA = ethanolamine (not an amino acid), Gln = glutamine, Glu = glutamic acid, Gly = glycine, His = histidine, Hse = homoserine (internal standard), Ile = isoleucine, Leu = leucine, Lys = lysine, Met = methionine, Orn = ornithine, Phe = phenylalanine, Ser = serine, Thr = threonine, Trp = tryptophan, Tyr = tyrosine. Val = valine. Note: Met peak was obscured by Trp in all samples except for five samples in Species 3.

and (d) L-asparagine, β-alanine, glycine, L-lysine, L-phenylalanine. The amino acids in each of these mixtures were well-separated. The standards were appropriately diluted to approximate the concentrations of the corresponding amino acids in the extract. The identified peaks corresponded to the 10 largest peaks and at least 15 of the 20 largest peaks in each species.

Data analysis

Chromatographic peak areas for each chromatogram were obtained by computer integration using commercial software (Computer Interface Consultants, Halifax, NS) and entered into a spreadsheet. These areas were then corrected for internal standard and dry mass. Prior to multivariate analysis, it was necessary to obtain accurate matching of peaks among samples. Great care was taken at this stage to ensure consistent identification. Samples run on a particular day were generally not difficult to match, but small changes in chromatographic conditions made day-to-day variations in retention times somewhat greater. Even so, matching of major and standard peaks was not a problem, since spiked samples were run each day, and relative retention times were used to identify minor peaks. This procedure resulted in 66 unique peaks (in addition to the internal standard, homoserine) which included identified and unidentified amino acids as well as primary amines also labelled by the derivatization. A statistical summary of corrected peak areas for these peaks is given in Table 2. For each species, the table gives the mean value for each peak area, the relative standard deviation of the peak area, and the number of samples which exhibited that peak.

Multivariate data analysis was carried out on DOS compatible computers using in-house and commercial software. The latter included *Einsight* v. 3.0 (Infometrix, Seattle, WA).

RESULTS AND DISCUSSION

The mean concentrations of the calibrated amino acids in dried samples of each of the six species investigated are given in Table 3. The sum of these amino acid concentrations was similar for each species except for *A. virosa*, which

TABLE 3

Mean concentrations of amino acids in *Amanita* species (Relative standard deviations are the same as in Table 2)

Amino acid	Mean concentration by species (μmol per g of dried mushroom [a])						
	Flavorubescens	*Virosa*	*Flavoconia*	*Rubescens*	*Gemmata*	*Muscaria*	Mean
Asp	6.7	1.6	3.7	6.6	8.6	9.4	6.1
Glu	10.6	3.3	12.3	12.3	13.0	23.3	12.5
Asn	10.7	1.8	9.1	10.2	6.8	9.6	8.0
Ser	14.3	6.5	7.8	10.2	12.6	14.6	11.0
Gln	39.1	7.5	25.2	37.0	36.5	27.2	28.8
His	15.9	5.2	21.8	20.5	9.2	10.5	13.9
Gly	9.2	5.7	4.0	20.4	5.4	6.5	8.5
Thr	5.5	5.1	7.4	5.6	5.4	6.4	5.9
Arg	20.6	6.8	9.0	13.4	25.7	23.2	16.4
Ala	48.5	22.8	22.1	40.4	23.8	23.8	30.2
Tyr	0.4	0.6	4.0	0.1	1.9	0.1	1.2
Trp/Met [b]	1.7	0.6	1.5	0.4	4.6	1.4	1.7
Phe	1.0	0.0	3.5	0.8	5.5	1.7	2.1
Ile	6.6	1.5	2.0	3.3	1.2	2.1	2.8
Leu	3.3	1.2	2.8	3.7	4.0	4.1	3.2
Total	194.1	70.2	136.2	184.9	164.2	163.9	152.3

[a] Mean moisture content = (90 ± 4)%. [b] Trp and Met were not fully resolved and were calibrated as tryptophan only.

contained less than one-half of the mean value for the six species. The values for these *Amanita* samples were less than 30% of the concentrations for the same set of amino acids reported for *Agaricus bisporus* [19] and *Pleurotus ostreatus* [20]. Averaged over the six species, alanine, glutamine, arginine and histidine are the four most abundant amino acids and correspond to four of the six most abundant identified amino acids in each species. The predominance of free alanine in the fungi is consistent with the other work [10,15, 19,20] and high concentrations of glutamine [10,19,20] and arginine [15,19] seem to be common. High concentrations of histidine appear to be less prevalent in other fungi. The distribution of amino acids within the six *Amanita* species appeared similar and did not permit an immediate distinction among species.

The initial step in the exploratory data analysis was the selection of features for pattern recognition, since the number of peaks given in Table 2 was far to large to allow them to be used directly. A total of 66 different peaks were found in the six species, and the number of peaks found in a single species ranged from 38 to 50. These peaks were classified according to their occurrence across samples and species and the results are summarized in Table 4. The classification of groups is as follows: (A) peaks which are present in all samples, (B) peaks which are observed for all species, but not all samples, (C) peaks found in all samples of a limited number of species (i.e., 1 to 5), (D) peaks absent from at least one species, but present in some samples of at least one other species and all samples of at least one other species, and (E) peaks found in a limited number of species (i.e., 1 to 5) but not in all samples of any species. It can be seen from the table that 30 peaks (Groups A and B) were present in all species and that 22 of these were present in all samples analyzed. Peaks which are potentially more useful for direct classification are those that do not occur in all species (Groups C, D, and E) and previous chemotaxonomic investigations have focused on the limited distribution of individual components including non-protein amino acids in fungal species [9,14,21]. Of these three groups, peaks in Groups D and E do not occur in all samples and therefore cannot be regarded as completely reliable for classification. The most useful peaks for this purpose are those present in all samples of one (or a few) species (Group C). Only three peaks are found in this group. Of these, two are common to two species (peaks 44 and 65), and one is found in five of the six species (peak 41). Peak 44 was found in *A. gemmata* and *A. muscaria* and has been identified through spiking as the γ-aminobutyric acid (GABA) agonist muscimol, a known constituent of these species [9,21]. The other peaks in Group C remain unidentified. In any case, there is no single peak or set of peaks that can be used to unambiguously and reliably identify a particular species.

For the purposes of pattern recognition studies, it was decided that peaks that did not occur in all samples could not be considered reliable as features. Many of these were very small and all were unidentified. This limited the possible choices to Group A, which contained most of the largest peaks (17 of 20) and all of the identified

TABLE 4

Summary of peaks by species and pattern of occurence

Group	*Amanita* species						All species
	Flavorubescens	*Virosa*	*Flavoconia*	*Rubescens*	*Gemmata*	*Muscaria*	
A	22	22	22	22	22	22	22
B	8	8	8	8	8	8	8
C	2	1	0	2	2	2	3
D	6	5	7	4	3	4	8
E	12	9	8	3	3	2	25
Total	50	45	45	39	38	38	66

peaks. Of the 20 identified peaks in this group (the internal standard, Hse, was not included in this count), tryptophan and methionine were excluded since they were not fully resolved in most samples. Ethanolamine was excluded as it was not an amino acid. The unidentified peaks were also excluded since their origins were uncertain. This left 17 amino acids that formed a base data set of potential features. Subsets of these were further extracted as described below.

The next step in the data analysis involved an examination of correlations present. At this stage, all 66 peaks present in the original data set were used for a thorough comparison. Correlation coefficients for each peak with every other peak were computed over all 59 samples and very few significant correlations were found. Only five pairs exhibited correlations greater than 0.8, and of these the only case where both peaks were identified was for asparagine and glutamine. These are the amide forms of aspartic and glutamic acid, respectively, and glutamine, a key intermediate in nitrogen metabolism, serves as the nitrogen donor for the synthesis of asparagine from aspartic acid. Their high correlation (0.865) may be related to the availability of nitrogen, which is often an important factor in plant metabolism [15]. All but a few of the correlations were positive, reflecting the fact that the data set was not normalized for the total amino acid content and an increase in one component generally meant an increase in the others.

The correlations among species were also examined. This was accomplished by calculating correlation coefficients for each sample in one species against each sample in another species using all peaks and then averaging over all pairings. The results are shown in Table 5. Note that the diagonal elements of the correlation matrix are not unity because they represent the average of sample correlations within a species. In general, the correlations are fairly high. As expected, intra-species correlations are greater than inter-species correlations with two exceptions, *A. flavoconia* and *A. rubescens*. Both of these gave high correlations with *A. flavorubescens*. In the case of *A. flavoconia*, this is not unexpected since there is also a strong morphological similarity with *A. flavorubescens* [7].

Initial efforts to separate all six species in two-dimensional principal components space were unsuccessful, so the problem was limited to three species which showed good promise of separability, *A. flavoconia*, *A. gemmata* and *A. virosa*. The first step involved feature selection through the calculation of mean Fisher weights [2], which is the ratio of the square of the difference of means of a feature for two classes to the sum of the variances. Only the identified amino acids previously noted were included in this calculation. The corrected areas for six amino acids were selected: His (6.32), Leu (5.15), Phe (4.16), Glu (3.70), Tyr (3.37) and Asn (3.36) (mean Fisher weights are given in parentheses). The number of features was limited to six since the use of additional amino acids did not significantly affect the species separation. Exploratory analysis with this set produced a good separation by both principal components and hierarchical clustering, with the first two principal components accounting for 86.5% of the variance. The dendrogram, constructed using the incremental sum of squares method [22] with autoscaled values, is shown in Fig. 3. Only one sample was misclassified by this

TABLE 5

Correlations of chromatographic peaks among species

	Flavorubescens	*Virosa*	*Flavoconia*	*Rubescens*	*Gemmata*	*Muscaria*
Flavorubescens	0.930	0.833	0.784	0.857	0.833	0.771
Virosa		0.943	0.653	0.762	0.633	0.641
Flavoconia			0.774	0.745	0.715	0.697
Rubescens				0.779	0.749	0.712
Gemmata					0.870	0.802
Muscaria						0.854

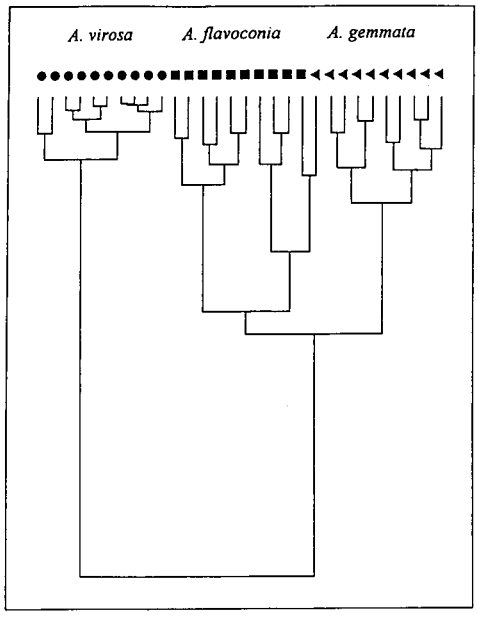


Fig. 3. Dendrogram of three species separated using corrected areas for the six best amino acids.

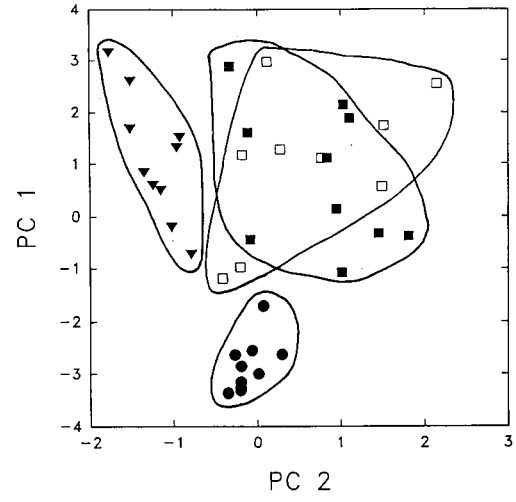


Fig. 4. Scatter plot of four species in principal components space generated from corrected peak areas.

method. The addition of the species *A. flavorubescens* to the data set resulted in the selection of a slighlty different set of six amino acids: His (3.72), Leu (3.27), Phe (3.21), Asn (2.89), Glu (2.83) and Gln (2.62). The principal components plot in Fig. 4 shows that the initial three species remain well separated, but there is strong overlap of *A. flavoconia* and *A. flavorubescens*, which again may be related to their morphological similarity. Addition of the other two species to this space produced a broad overlap.

It was postulated that the limited success of the initial exploratory analysis might be due to variations in the total amino acid content and that changes in the distribution of amino acids might be more indicative of species. Cohen and Farnham [15] have shown that the total amino acid content of certain fungi vary with the developmental stage, tending to decrease with age. To circumvent this problem, ratios of corrected peak areas to other selected peaks in the chromatograms were used as features. Although a variety of peaks were selected for this normalization, none produced results that were significantly superior to those obtained by using the corrected peak areas directly. This is possibly because any advantage gained may be neutralized by the larger feature variances resulting from the calculation. As an alternative, the total peak area of each chromatogram was used as a normalization factor for each data set. In doing this, it was assumed that the detector sensitivity was the same for each fluorescent derivative. While calibration data are not available for all peaks, relative sensitivity factors given in Table 1 indicate that this is a reasonable approximation. Treatment of the data in this fashion yielded a better separation of the four species previously mentioned. The best separation was obtained with range scaling and rotation of the first three principal components, as shown in Fig. 5. There was also a significant change in the optimal descriptors, with the normalized data set giving Phe (10.1), Gly (5.42), Orn (2.94), Ala (2.17), Asn (2.04) and Leu (1.77) as the best amino acids. An especially large mean Fisher weight was found for phenylalanine and was found to be mainly due large differences between the species *A. gemmata* and two others, *A. virosa* and *A. flavorubescens*. A dendrogram constructed using the Lance and Williams flexible method [22], is shown in Fig. 6 and provided complete separation of all four species (no misclassifications). Addition of the two excluded species to the data set resulted in a somewhat

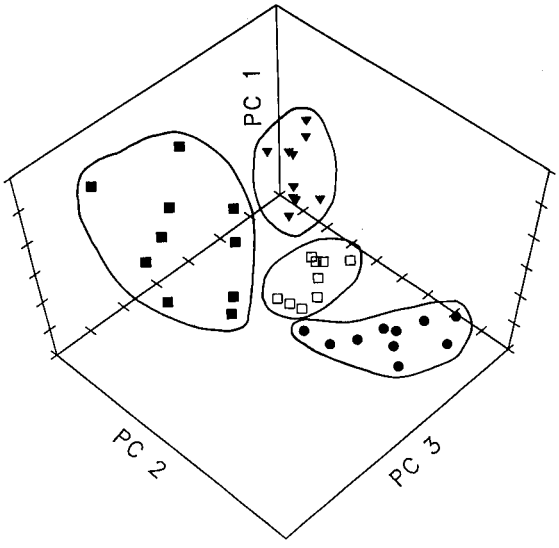


Fig. 5. Scatter plot of four species in principal components space (rotated) generated from normalized peak areas.

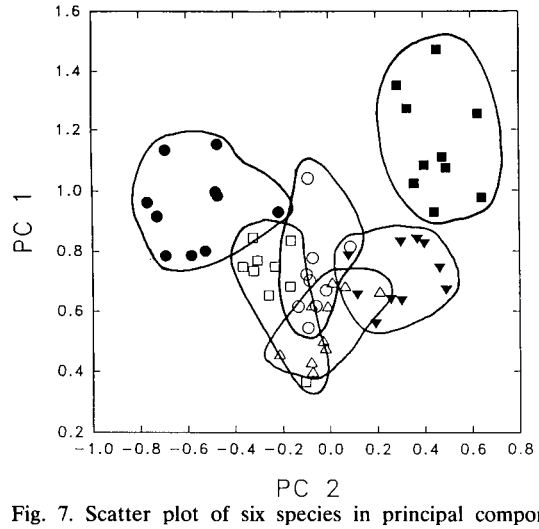


Fig. 7. Scatter plot of six species in principal components space generated from normalized peak areas.

modified set of descriptors [Phe (5.41), Gly (2.24), Ala (2.23), Orn (2.04), His (1.88) and Tyr (1.74)] but failed to produce a good separation of all species. The principal components plot for this case is shown in Fig. 7. Although complete class separation was not achieved, the additional species do form fairly tight clusters in the two-dimensional space, indicative of a cohesive data structure. The six species were also projected into the principal components space generated for Fig. 4, but this did not produce superior results. Dendrogram generation using all six species resulted in the misclassification of 17% of the samples. The method of k-nearest neighbours (KNN) [1,2] was also applied to the classification using the same six amino acids. The application of KNN with six species and $k = 1$, 3 and 5 produced correct classifications for 86, 83, and 79% of the samples, respectively.

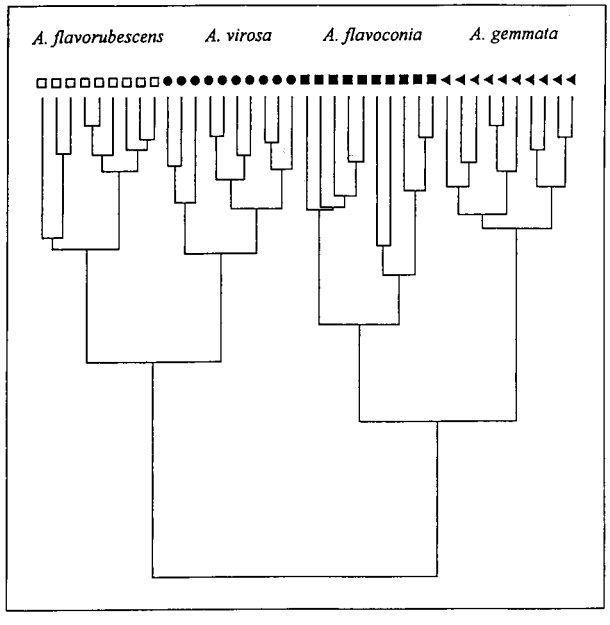


Fig. 6. Dendrogram of four species separated using normalized peak areas.

Since supervised pattern recognition involving feature selection can sometimes result in coincidental separation simply on the basis of random correlations in the descriptors, this possibility was evaluated by a Monte Carlo approach. The same data were used but classes (species) were assigned at random. Features were then selected for optimal class separation. None of these trials produced a significant separation of the randomly assigned classes. Therefore, the clusters observed for the true classes are considered to be significant.

From these results, it appears that the unam-

TABLE 6

Classification of amino acids according to biosynthetic origin (Underlined amino acids were the most useful for distinguishing species)

A	Glutamic acid, glutamine, ornithine, arginine, proline [a], lysine
B	Aspartic acid, asparagine, threonine, isoleucine, methionine
C	Alanine, leucine, valine
D	Serine, glycine, cysteine [b]
E	Phenylalanine, tyrosine, tryptophan
F	Histidine

[a] Not detected (amino group is secondary). [b] Not detected (cysteine does not form a fluorescent derivative with OPA).

biguous classification of *Amanita* fungi through the analysis of amino acid concentrations would be difficult, although some species, such as the deadly poisonous *A. virosa*, showed consistent separation from the other species in all cases. Routine classification would also be impractical because of the numerous species of mushrooms that can be found in a given region. For these reasons, further testing of classification abilities through the implementation of other algorithms and cross validation was not attempted. It does appear, however, that different species do exhibit significant differences in the distribution of free amino acids. Connections were sought between these differences and evolutionary or biochemical relationships among the fungi, but at present these are unclear. Table 6 lists six classes of amino acids based on common biosynthetic origins [23]. The position of each amino acid within a group represents its approximate position in the biosynthetic chain, although in some cases parallel and branched biosynthetic pathways exist. The underlined amino acids were found to be the most useful features in the results presented here. Although no clear patterns were observed, it was noted that in all four sets of results presented, the best features always represented five of the six amino acid classes, although not always the same five. Further relationships remain to be determined.

In summary, these results contradict some of the earlier studies that suggested no significant differences in amino acid content among fungi, although different species were investigated in most of those cases. In this work, more quantitative data have been examined and more rigorous methods have been employed. Except in special cases where the number of candidate species can be reduced, it is not likely that these differences can be used for classification purposes. Nevertheless, the class structure observed may be useful in identifying certain biological or biochemical relationships in these organisms.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors wish to thank K.L. Horton for her technical assistance, B.E. McCarry for providing a sample of muscimol, and D.L. Bunbury and A.L. Doherty for their contributions.

REFERENCES

1 D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988.
2 M.A. Sharaf, D.L. Illman and B.R. Kowalski, Chemometrics, Wiley, New York, 1986.
3 C.A. Stace, Plant Taxonomy and Biosystematics, Edward Arnold Publishers, London, 2nd edn., 1989.
4 V.H. Heywood and D.M. Moore (Eds.), Current Concepts in Plant Taxonomy, Academic Press, 1984.
5 P.M. Smith, The Chemotaxonomy of Plants, Edward Arnold Publishers, London, 1976.
6 B. Lavine and D. Carlson, Anal. Chem., 59 (1987) 468A.
7 O.K. Miller, Jr., Mushrooms of North America, E.P. Dutton, New York, 1984.
8 T. Wieland, Science, 159 (1968) 946.
9 W.S. Chilton and J. Ott, Lloyida, 39 (1976) 150.
10 R. Close, Nature, 185 (1960) 609.
11 S. Hatanaka and H. Terakawa, Bot. Mag., 81 (1968) 259.
12 P. Catalfomo and V.E. Tyler, Jr., J. Pharm. Sci., 50 (1961) 689.
13 J.E. Robbers, L.R. Brady and V.E. Tyler, Jr., Lloydia, 27 (1964) 192.
14 V.E. Tyler, Jr., R.G. Benedict and D.E. Stuntz, Lloydia, 28 (1965) 342.
15 R.J. Cohen and W.F. Farnham, Phytochemistry, 15 (1976) 251.
16 H. Faulstich, A. Buku, H. Bodenmuller and T. Wieland, Biochemistry, 19 (1980) 3334.
17 B.N. Jones and J.P. Gilligan, J. Chromatogr., 266 (1983) 471.

18 R.L. White, A.C. DeMarco, and K.C. Smith, J. Chromatogr., 483 (1989) 437.
19 Y. Oka, H. Tsuji, T. Ogawa and K. Sasaoka, J. Nutr. Sci. Vitaminol., 27 (1981) 253.
20 Y. Oka, T. Ogawa and K. Sasaoka, J. Nutr. Sci. Vitaminol., 30 (1984) 27.
21 R.G. Benedict, V.E. Tyler, Jr. and L.R. Brady, Lloydia, 29 (1966) 333.
22 Einsight User's Manual, Infometrix, Seattle, WA, 1991.
23 G. Zubay, Biochemistry, MacMillan, New York, 2nd edn., 1988, Chap. 24.

Analytica Chimica Acta, 277 (1993) 347–356
Elsevier Science Publishers B.V., Amsterdam

Applying and developing receptor models to the 1990 El Paso air data: a look at receptor modeling with uncharacterized sources and graphical diagnostics

Clifford H. Spiegelman and Stuart Dattner

Texas A. & M. University, College station, TX 77840 (USA)

(Received 3rd September 1992; revised manuscript received 23rd December 1992)

Abstract

This paper represents an ongoing receptor modeling research of airborne species in El Paso, Texas. It represents a six month collaboration between the authors. It extends the case study reported by Spiegelman and Dattner in 1992. For completeness the background material is reviewed.

Keywords: El Paso air data (1990); Graphical diagnostics; Receptor models

El Paso is the only city in Texas where humans have caused a violation of the standard for particulate matter of size less than 10 μm (PM10). El Paso shares a dry mountain valley with Ciudad Juarez, Chihuahua, Mexico. The pollution from both cities mixes freely. The combined population of these cities is between 1.5 and 2.0 million. During some winter periods, temperature inversions trap the air pollution in these cities. This, in part, is due to the fact that these cities share a mountain pass. Most air pollution episodes occur during the winter temperature inversions. The Texas State Air Control Board (TACB) has joined the US EPA, the El Paso City-County Health District (EPCCHD), Sandia National Laboratory, The Secretariat de Desarrolo Urbana y Ecologia de Mexico (SEDUE), and several US EPA contractors in a scoping study (our work is a small part of this larger study).

Correspondence to: C.H. Spiegelman, Texas A. & M. University, College Station, TX 77840 (USA).

SAMPLING METHODOLOGIES

Dichotomous PM10 samplers were operated at 5 ambient monitoring sites during this study. A dichotomous sampler collects particles in two size ranges. These two ranges were determined on the basis of potential health effects. The first size range is for course particles. It is for particles between 2.5 and 10 μm. The second size range is for fine particles. It is for particles less than 2.5 μm. The course particles are usually collected in the upper respiratory tract. The fine particles are collected in the human lung and are potentially more toxic than course particles. The five sites are shown in Fig. 1. Two of these sites, SUN and CHM, were chosen because they were located within several hundred feet of the Texas–Chihuahua border. Another two sites ADV and TEC were located in Juarez and were used to determine concentrations in that city. The last site, CM6, was located near the major US highway (Interstate 10) and downtown El Paso. This site was chosen to be representative of the city of El Paso.

The samples collected in this paper were collected on clean preferred, quartz filters. The samples were analyzed by thermal decomposition techniques. The general categories for our species measured are organic carbon (tar), elemental carbon (soot), and inorganic carbon [1]. The samples were kept refrigerated after collection to minimize the loss of the organic carbon. The technique used by TACB laboratory produces a set of observations that include: (1) five volatility based fractions of organic carbon; (2) an elemental carbon value, and (3) a carbonate value; (4) other variants of the above.

The variables listed above can be considered as different. In addition the course and fine measurements can also be considered as different (from a linear algebraist point of view they are independent). The 15 species that we used are in both fine and coarse species: low temperature organics, medium temperature organics, high temperature organics and inorganic carbonate, charred organics and low temperature elemental carbon fraction, high temperature elemental carbon, elemental carbon, organic carbon. In addition we used coarse inorganic carbonate (fine inorganic carbonate was usually below the detection limit). Theoretically not all of these species are linearly independent but due to missing species on some days (and measurement errors) all 15 variables were used. Two discrete 12-h collection periods were used each day. Roughly they went between 5 a.m. and 5 p.m. There were missing data. The data that we used are the percent of mass for each of 15 species. They are

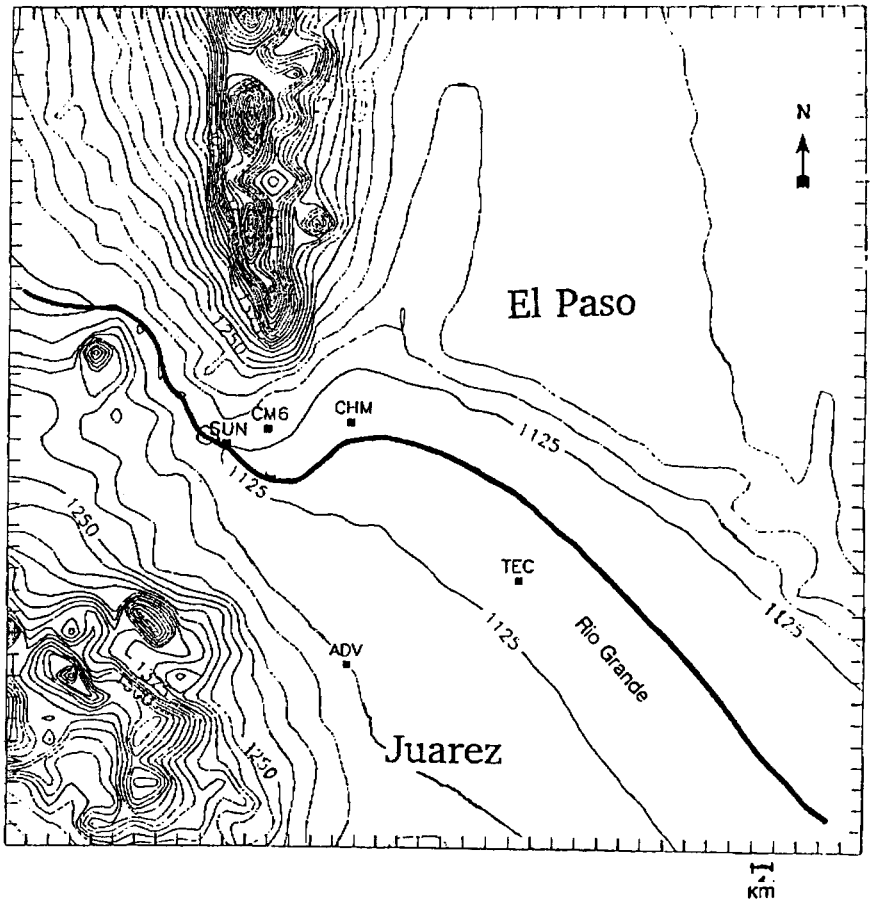


Fig. 1. The 5 ambient monitoring sites where dichotomous PM10 samplers were operated.

called mass fraction spectra and have no deep link to Fourier transforms or time series analysis. (Fourier transformations, called spectra, are a common statistical tool. Statisticians call mass fraction spectra probability mass functions. However, the mass functions are for chemical species not values of real valued random variables.) Each collected spectrum is the convex combination of spectra from different pollution sources. We call them pure sources. Each pure source, e.g., automobile emissions, has its own spectra and it is hoped that different types of sources have different spectra. The main problem that receptor modelers face is decomposing observed spectra into contributions from source spectra. For example 10% of the observed spectra may be due to auto emissions, 20% to coal burning power plants, 25% wood burning stoves, etc. Many receptor modelers have knowledge about source spectra shape and/or the number of spectra. However, sometimes this information is unavailable. This is the case in our work. Thus we have to estimate these pure source spectra from the data as well as the number of sources. In addition we give a minimax strategy that helps to conservatively compensate for missing pure source spectra. We will also estimate the mean source locations for many sources. Other than Spiegelman and Dattner [2], we have not previously seen receptor models used (in our context) to estimate mean locations for sources of pollution. Finally we will also formulate an approach for locating places to put new collectors.

BACKGROUND FOR OUR RECEPTOR MODELING APPROACH

We focus on our implementation of receptor modeling, other approaches were mentioned by Spiegelman and Dattner [2]. We first considered the models and methods developed by leading chemometricians [3–8].

After looking at these models we decided that existing receptor methods would be modified for the data set on hand. However, many of ideas from existing receptor models are used. In particular in our case both the number of pollution sources and the pure source mass fraction spectra are unknown. Henry and Kim [6] and Henry et al. [8] insisted that mass fraction spectra be feasible. That is mass fraction spectra must make physical sense. This is a very important idea and we borrow it. However, their method did not pick a finite set of mass fraction spectra to use as contribution source mass fraction spectra.

Model

The basic model is:

$$Y_i = \sum_{k=1}^{q} \alpha_{ik} P_k + \epsilon_i, \quad \text{for } i = 1, \cdots, n \tag{1}$$

Here the Y_i are observed mass fraction spectra (collected at n time periods) and the P_k are pure source mass fraction spectra (source contributions). The Y_i values are observed, the P_k values are not. The α_{ik} are the weights, that satisfy the following constraints $\Sigma_{k=1}^{q} \alpha_{ik} \leq 1$, $\alpha_{ik} \geq 0$. The number of pure sources q is also unknown. Note that unlike principal component analysis the number of pure sources may be greater than the number of species. The ϵ_i values also contain source mass fraction spectra, and correspond to unmodeled mass fraction spectra and normalized measurement error. The errors may not have mean 0.

Literature approaches

Hopke [3] fitted known mass fraction spectra P_k to the observed mass fraction spectra Y_i using the constraints on the α_{ik}. Henry and Kim [6] used a singular valued decomposition of observed mass fraction spectra to find feasible contribution source mass fraction spectra. In Ref. 9 it is assumed that one pure source mass fraction spectrum is known and one is unknown. They make an assumption that in the limit the unknown source is observed. We will use a related idea below. They used Dirichlet priors on the unknown α values. In an unpublished work Bandeen-Roache [10] discusses identifiably conditions for source mass fraction spectra. We take the following approach in order to estimate some of the unknown mass fraction spectra.

Methods

(1) We collapse the observed spectra to two dimensions. That is $P_i = (p_{1i}, \cdots, p_{15i})$ and we replace this P_i by 15 two-dimensional spectra $(p_{ji}, 1 - p_{ji})$. These two dimensional spectra lie on the line with (generic) equation $y = 1 - x$. For example one of these two-dimensional spectra represents the mass fraction of coarse organic carbon and the mass fraction of everything else (used). On any fixed collection period (morning or night) we considered the ratios of the two-dimensional (one-parameter) spectra. The data that we have is X_i. These X_i measurements are related to the Y_i by the following equation:

$$X_i = (x_{i1}, \ldots, x_{i15})$$
$$Y_i = \frac{X_i}{\sum_{j=1}^{15} x_{ij}} = (y_{i1}, \ldots, y_{i15}) \qquad (2)$$

If a ratio was greater than 10 we considered these spectra as candidates for near pure source spectra. The ratios of the collapsed spectra at a fixed species are used to search for almost pure source spectra. Suppose PS_1 and PS_2 are the two extremes among pure sources for a specie. Further suppose that $PS_1 \geq PS_2$. For this species each spectrum is a convex combination of PS_1 and PS_2. Our criteria for a potential pure source are

$$\frac{\alpha_1 PS_1 + (1 - \alpha_1)PS_2}{\alpha_2 PS_1 + (1 - \alpha_2)PS_2} \geq 10$$

Result: if $[\alpha_1 PS_1 + (1 - \alpha_1)PS_2]/[\alpha_2 PS_1 + (1 - \alpha_2)PS_2] \geq 10$ then $\alpha_1 \geq 10\alpha_2$ or $\alpha_2 \leq 1$.

Proof: see Ref. 2.

In addition graphical techniques were used to locate novel spectra shapes. The data set was big enough that some criteria like the one given above was needed to reduce the number of spectra graphed.

The idea that we use is that every point on the line segment can be written as a convex combination of the line endpoints. Theoretically the endpoints are from pure source two dimensional spectra. (Note that several pure source spectra may lead to the same two dimensional spectra.) We explore this in the next section. We know that this procedure will miss some pure spectra. It is easy to construct examples. We ignore the fact that some pure spectra that we identify may be due to random error. This is because we hope that the random errors are small.

(2) We fit the observed spectra to the estimated pure spectra using the model:

$$\begin{aligned} &\underset{\alpha_{ik}}{\text{MAX}} \sum_{k=1}^{q} \alpha_{ik} \\ &Y_{i(1)} \geq \sum_{k=1}^{q} \alpha_{ik} P_{k(1)} \\ &\sum_{k=1}^{q} \alpha_{ik} \leq 1 \\ &\alpha_{ik} \geq 0 \end{aligned} \qquad (3)$$

The mass fraction spectra subscript in brackets indicates that the inequality is for each species. For every non zero $\hat{\alpha}_{ik}$ there is a tight constraint, i.e., some 1 such that $Y_{i(1)} = \Sigma_{k=1}^{q} \alpha_{ik} P_{k(1)}$. Thus if the errors in Y_i are small and the non-tight constraints are well separated then the distribution of the $\hat{\alpha}_{ik}$ can be gotten as a linear transform of the Y_i. If not then a complicated mixture distribution results. We did a small Monte Carlo study. We used the estimated pure sources and multiplied each component of the pure sources by an independent uniform random variable on the interval [0.7,1.3]. This simulation corresponds to $\pm 30\%$ error in the spectra which is what we feel the science dictates. The pure spectra were renormalized and a linear programming fit performed. The results show that the individual $\hat{\alpha}_{ik}$ are highly variable but the $\Sigma_{k=1}^{q} \hat{\alpha}_{ik}$ is stable. For example individual $\hat{\alpha}_{ik}$ values have standard deviations of about 0.1 while $\Sigma_{k=1}^{q} \hat{\alpha}_{ik}$ has a standard deviation of about 0.02. It is common for individual $\hat{\alpha}_{ik}$ values to range between 0 and 0.25, while the sum varies between 0.79 and 0.84 for a set of 5 simulation runs. It is our conjecture that this variation is do, in large part, to the highly multicolinear pure source spectra.

Since there is a great chance that some pure source spectra are missing in our fit we interpret $\text{MAX}_{\alpha_{ik}} \Sigma_{k=1}^{q} \alpha_{ik}$ as variation on a type 1 (or type 2) (from ANOVA)R^2. It is between zero and one. The expression $\Sigma_{k=1}^{q} \hat{\alpha}_{ik}$ needs to be interpreted.

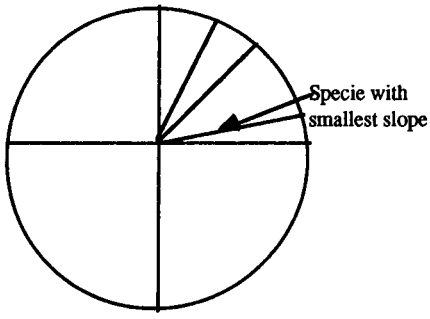


Fig. 2. Graph showing the $0 \leq \Sigma_{k=1}^{q} \hat{\alpha}_{ik} \leq 1$ relationship.

Let $\hat{\alpha}_{ik}/\Sigma_{k=1}^{q}\hat{\alpha}_{ik}$ be denoted as $\hat{B}_{ik}$. Then we have:

$$\sum_{k=1}^{q} \hat{\alpha}_{ik} = \underset{1}{\text{MIN}} \left[\frac{Y_{i(1)}}{\sum_{k=1}^{12} \hat{B}_{ik} P_{k(1)}} \right] \tag{4}$$

Since the mass fraction spectra are probability mass functions $0 \leq \Sigma_{k=1}^{q}\hat{\alpha}_{ik} \leq 1$ (The case when all the $\hat{\alpha}_{ik} = 0$ is trivial.) A graph showing this relationship is shown in Fig. 2.

Later we discuss a minimax strategy that takes into account missing spectra.

(3) We provide an estimate of a lower bound for the number of missing spectra. That is we provide an estimate that is biased low for the number of pure source spectra that would be identified by tactic one. Since (1) misses some pure spectra there is a further bias. It is easy to construct artificial spectra that our identification method misses. Still it is interesting to know how many more pure sources we might find if we were to collect more data. Estimating the number of species is an old problem in genetics and has come up in many different contexts. For example estimating the number of words that Shakespeare knew is the subject of Efron and Thisted [11]. A recent paper discussing some of these issues is published by Loh [12]. We choose our own estimator here because:

(1) We are interested in the total number of species. The estimators in Efron and Thisted [11] are not designed to answer this question. The estimators are more suitable for answering how many pure sources we might encounter if we double the collection period.

(2) In our case the estimates in Efron and Thisted [11] provide with too high positive probability estimates that might give a negative number of new pure source mass fraction spectra. However for our example their estimates do not give a negative number of new pure species, but slight changes in what we call a repeated occurrence of a new species could produce this result.

We use the following estimator: (number of pure spectra observed)^2/(expected number that would be observed in the same time period if the observed pure spectra represented the whole population). The intuition behind this estimator is given by the following equation and estimates:

$$\frac{\text{number of observed pure spectra}}{\text{mean number to appear in similar time frame given \# of observed pure spectra}} = \frac{\text{true number of pure spectra}}{\text{mean number to appear in similar time frame given true \# of pure spectra}}$$

The data we used for the species calculation are:

1	2	3	5	6	7	11	13
4	1	1	2	1	1	1	1

The top row gives the number of times a pure source occurred and the bottom row gives the number of pure sources in the category above. For example 4 pure sources occurred once while one pure source occurred 13 times. For this data our integer estimate of the number of remaining pure sources that would not reoccur is 2. A naive confidence interval ignoring the uncertainty in estimated binomial probabilities and using ± 3 sigma (keeping estimates non-negative) is [0,5] new species. We note that the usual plug in binomial estimates of uncertainty may not work well because they involve moments of order 70. For the data in the table above the first estimator given in Efron and Thisted [11] gives 8 estimated new pure spectra. That is 8 new pure spectra in another 18 day collection period.

While our estimate is guaranteed to be non-negative it is clearly biased low. Further statistical

details are reported by Spiegelman and Dattner [2].

(3) We give a minimax strategy that conservatively compensates for missing pure source spectra. The objective function $\mathrm{MAX}_{\alpha_{ik}} \Sigma_{k=1}^{q} \alpha_{ik}$ is a quasiconcave function of the mass fraction spectra. What we propose as a conservative strategy is adding a mass fraction spectra to the fit so as to minimize the $\mathrm{MAX}_{\alpha_{ik}} \Sigma_{k=1}^{q} \alpha_{ik}$. More specifically we propose solving the following:

$$\begin{aligned} &\underset{P^*}{\mathrm{MIN}}\ \underset{\alpha_{ik},\alpha^*}{\mathrm{MAX}} \left(\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \right) \\ &Y_{i(1)} \geq \sum_{k=1}^{q} \alpha_{ik} P_{k(1)} + \alpha^* P^* \\ &\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \leq 1 \\ &\alpha_{ik} \geq 0,\ \alpha^* \geq 0 \end{aligned} \tag{5}$$

Let P_1^* and P_2^* be two source mass fraction spectra and $(\Sigma_{k=1}^{q} \alpha_{ik}) + \alpha_1^*$ and $(\Sigma_{k=1}^{q} \alpha_{ik}) + \alpha_2^*$ be the corresponding solutions to Eqn. 5.

Lemma. If α_{ik} are fixed in Eqn. 5 and not maximized then α^* is quasiconcave functions of the P^* and given α^* (treated as fixed) the solutions $\Sigma_{k=1}^{q} \alpha_{ik}$ are a concave function of P^*. In other words α^* in

$$\begin{aligned} &\underset{P^*}{\mathrm{MIN}}\ \underset{\alpha^*}{\mathrm{MAX}} \left(\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \right) \\ &Y_{i(1)} \geq \sum_{k=1}^{q} \alpha_{ik} P_{k(1)} + \alpha^* P^* \\ &\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \geq 1 \\ &\alpha_{ik} \geq 0,\ \alpha^* \geq 0, \text{ and } \alpha_{ik} \text{ fixed} \end{aligned}$$

is a quasiconcave function of P^*. The $\Sigma_{k=1}^{q} \alpha_{ik}$ in

$$\begin{aligned} &\underset{P^*}{\mathrm{MIN}}\ \underset{\alpha_{ik}}{\mathrm{MAX}} \left(\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \right) \\ &Y_{i(1)} \geq \sum_{k=1}^{q} \alpha_{ik} P_{k(1)} + \alpha^* P^* \end{aligned}$$

is a concave function of P*.

$$\begin{aligned} &\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \leq 1 \\ &\alpha_{ik} \geq 0,\ \alpha^* \geq 0, \text{ and } \alpha^* \text{ fixed.} \end{aligned}$$

Proof. We show the proof for the first part only. The proof of the second part is similar. Consider the spectra $\lambda P_1^* + (1-\lambda) P_2^*$. It is clear that

$$Y_{i(1)} \geq \sum_{k=1}^{q} \alpha_{ik} P_{k(1)} + \lambda \alpha_1^* P_1^* + (1-\lambda) \alpha_2^* P_2^*.$$

Then

$$\begin{aligned} Y_{i(1)} \geq &\sum_{k=1}^{q} \alpha_{ik} P_{k(1)} + \mathrm{Min}(\alpha_1^*, \alpha_2^*) \\ &\times \left(\lambda P_1^* + (1-\lambda) P_2^*\right). \end{aligned}$$

It follows from the proof of the lemma above that if we consider the following problem:

$$\begin{aligned} &\underset{\alpha^* P^*}{\mathrm{MIN}}\ \underset{\alpha_{ik}}{\mathrm{MAX}} \sum_{k=1}^{q} \alpha_{ik} \\ &Y_{i(1)} \geq \sum_{k=1}^{q} \alpha_{ik} P_{k(1)} + \alpha^* P^* \end{aligned}$$

the solution $(\Sigma_{k=1}^{q} \alpha_{ik})^*$ is a concave function

$$\begin{aligned} &\sum_{k=1}^{q} \alpha_{ik} + \alpha^* \leq 1 \\ &\alpha_{ik} \geq 0,\ \alpha^* \geq 0 \end{aligned}$$

of $\alpha^* P^*$. This will allow rapid calculation of near optimal solutions of Eqn. 5.

(4) We provide a map of the mean location of the pure sources that we found. (For example the mean location for auto pollution would be a center of mass calculation weight given to the amount and location of each automobile.) As a part of this scoping study, we give a method for determining whether an important part of El Paso's air pollution data is imported from Mexico. Since we are working with mass fractions we cannot determine absolute magnitudes. However, if we assume a primitive transport model we can give estimates of the mean locations of the pollutants.

Assumption. In model 1 we assume that on collection period i that the mean distance of the kth source from the jth collector is proportional to α_{ik} and that these constants are unique. (Note that linear programming solutions are not always unique. If they are not, what follows would have to be done for the set of solutions.) We further

assume that the proportionality constant is the same for all collection sites. This assumption implies that El Paso has no wind and that pollutants settle out proportionally with distance to collectors. If the wind blows from Juarez to El Paso then the location estimates will be biased toward El Paso. A final assumption is that the mean location of all pollution sources lies within the convex hull of the collection site locations.

Let the location of the jth collector be given as (u_j, v_j), $j = 1, \cdots, 5$. Under the assumption above the mean location of the kth pollution source on the ith day is

$$\frac{\sum_{j=1}^{5} \alpha_{jik}(u_j, v_j)}{\sum_{j=1}^{5} \alpha_{jik}}$$

Since we do not know the α_{jik} we use the estimate

$$\frac{\sum_{j=1}^{5} \hat{\alpha}_{jik}(u_j, v_j)}{\sum_{j=1}^{5} \hat{\alpha}_{jik}}$$

As stated before the estimates of α_{jik} are highly variable. Figure 3 gives the locations of the collection sites and the estimated mean location of the pollution sources. A second simulation run was done by adding uniform error of $\pm 30\%$ to the pure source mass fraction spectra. The results of this simulation are shown overlaid with the

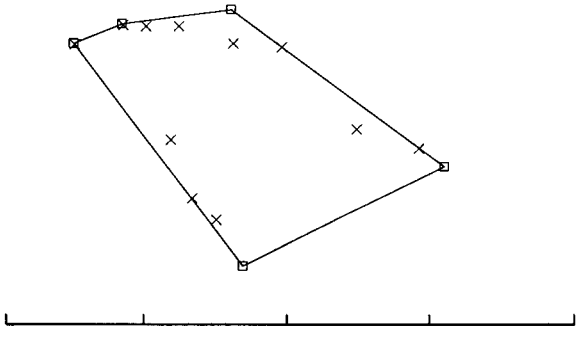


Fig. 3. Locations of collection sites and estimated mean location of the pollution sources.

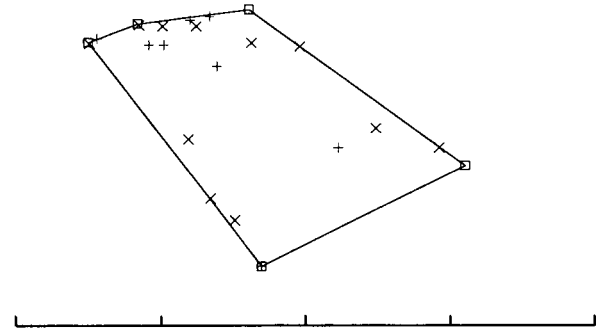


Fig. 4. Simulation results.

initial estimates (see Fig. 4). Fewer pure sources showed in the simulation sources although it is possible to argue that the general pattern of pollution source locations stayed the same. However, it is hoped that by averaging over several days that we can reduce some of the variability without introducing extra sources of variability such as changing wind directions. The effect of meteorological, and realistic physical science models on our estimates is under investigation. In later studies we hope to tag fixed location and variable location sources. Thus the automobile mean source location is expected to be more variable than a mean power plant location. By studying the variation in the $\hat{\alpha}_{ik}$ we hope to further determine the mobility and multiplicity of the sources.

(5) We provide an algorithm for placing future collection sites. The sites will provide a compromise choice between covering a larger geographic area and lowering the uncertainty in mean locations with the existing area. An important issue for pollution monitoring is the placement of monitoring sites. There are many issues to consider: estimation of pollution source locations, detection of polluters, regions without polluters, etc. We give a criterium for the placement of new collection sites that uses a strategy described in Ref. 13. Ours is an incremental approach in the same spirit found in Ref. 14. Due to weather conditions, terrain, and other factors, the further a collection site is from the pollution source the less accurate are determinations from that site.

We choose future monitoring sites to minimize the mean squared error of the estimator

$$\frac{\sum_{j=1}^{5} \hat{\alpha}_{jik}(u_j, v_j)}{\sum_{j=1}^{5} \hat{\alpha}_{jik}}$$

We assume both the bias and variance of this estimator depend upon the distance of the sites from the source. In particular we assume that:

$$\left| E\frac{\hat{\alpha}_{jik}}{\sum_j \hat{\alpha}_{jik}} - \frac{\alpha_{jik}}{\sum_j \alpha_{jik}} \right| \leq M_i \| S - \omega_j \|$$

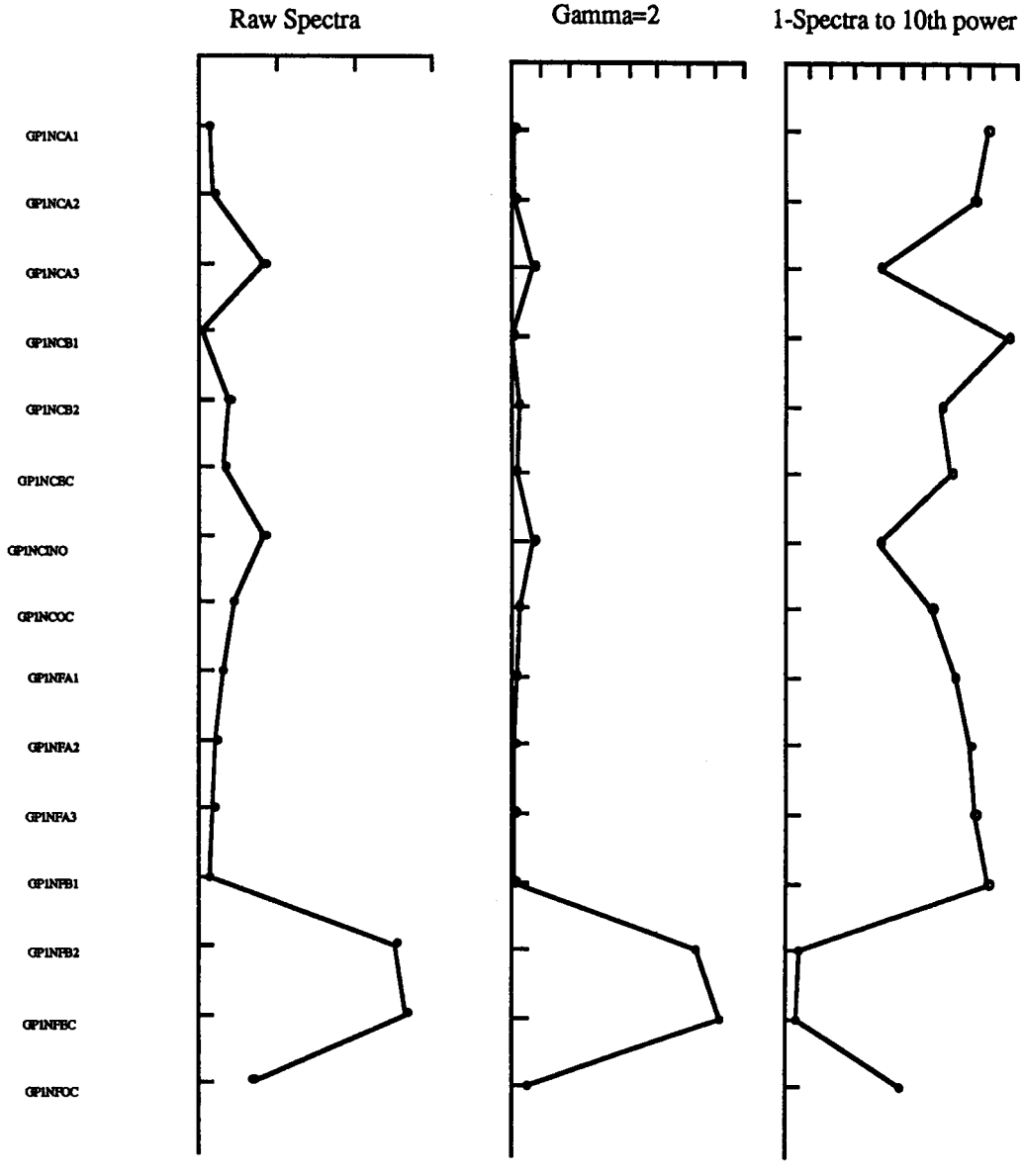


Fig. 5. Plot of untransformed vs. transformed spectra.

and in addition we assume that the $(\hat{\alpha}_{jik})/(\Sigma_j \hat{\alpha}_{jik})$ have a Dirrichlet distribution with parameters $v_j = \sigma_j^2[A - \| S - \omega_j \|]^2$. (We pay little attention to the subscripts i and k here.) Thus the bias is bound by $\Sigma_{j=1}^5 M_j \| S - \omega_j \| (|u_j|, |v_j|)$ and the variance of

$$\frac{\sum_{j=1}^{5} \hat{\alpha}_{jik}(u_j, v_j)}{\sum_{j=1}^{5} \hat{\alpha}_{jik}}$$

for the first component is

$$\frac{\sum_j v_j \left[\left(\sum_i v_i \right) - v_j \right] u_j^2 + \sum_{j,r} v_j v_r u_j u_r}{\left(\sum_i v_i \right)^2 \left(\sum_i v_i + 1 \right)}$$

The second component has a corresponding expression for variance. Here S is the source location, $\omega_j - (u_j, v_j)$ is the location of jth collection site, A is a constant chosen to reflect bigger variances close to the collection site. (It is common in chemical measurements to have constant coefficient of variation.) We choose both the latitude and longitude for each collection site to minimize the integrated expected mean square error. The integration will assume that the source location S is a random variable uniformly distributed within the convex hull of the existing collection sites. This uniform distribution is chosen as a for instance. In the same spirit $\sigma_j^2 = \sigma^2$ and $M_j = M$.

(6) A plotting strategy that will help detect pure source spectra. The selection method that we use chooses spectra that have high or low values for individual species. Let $P_i^\gamma = (p_{1i}^\gamma, \cdots, p_{15i}^\gamma)$ and $(1 - P_i)^\gamma = [(1 - p_{1i})^\gamma, \cdots, (1 - p_{15i})^\gamma]$. We plot the first of these spectra for gamma between 2 and 3, and the second for gammas between 10 and 15. These plots are similar to those obtained by using maximum entropy signal enhancement [15]. In Fig. 5 we plot the untransformed spectra along with the two transformed spectra, P^2 and $(1 - P)^{10}$.

There are many other issues that need to be addressed. How many species measurements should be used to form a spectra? If the number of species is large then irrelevant species might determine the fit. If the number is too small we might miss an important specie. Can the variation in the estimated locations over time be used to separate fixed sources (power plants) from mobile sources (automobiles)? How can the estimate of pure source locations take into account variance information, weather, etc. These and many other issues will be addressed in future work.

IMPLEMENTATION AND FURTHER DETAILS

Many details that were provided by Spiegelman and Dattner [2] are skipped here.

If $PS_2 \neq 0$ the inequality can be improved.

Thus with a spectra having $\alpha_2 \approx 0$ we have candidate pure spectra. We also graph both spectra if the ratio is ≥ 10 to see if it looked unique or if our identification method misses anything obvious.

We found 12 spectra to use as estimated pure sources.

This work was supported in part by the TACB and the National Science Foundation. We thank Karen Bandeen-Roache for making a preprint available and Phil Hopke for helpful phone conversations.

REFERENCES

1 R. Cary, Carbon Aerosol Analysis Organic/Elemental/Carbonate, Report to the TACB, Sunset Laboratories, Forest Grove, OR, 1991.
2 C.H. Spiegelman and S. Dattner, Multivariate Chemometrics a Case study: Applying and Developing Receptor Models for the 1990 El Paso Winter PM10 Receptor Modeling Scoping Study in G.P. Patil and C.R. Rao (Eds.), Multivariate Environmental Statistics, Elsevier, Amsterdam, in press.
3 P. Hopke, Receptor Modeling in Environmental Chemistry, Wiley, New York, 1985.
4 P. Hopke, Chemom. Intell. Lab. Syst., 10 (1990) 21.
5 P. Hopke, Receptor Modeling for Air Quality Management, Elsevier, Amsterdam, 1991.
6 R.C. Henry and B.M. Kim, Chemom. Intell. Lab. Syst., 8 (1990) 205.
7 R.C. Henry, Atmos. Environ., 21 (1987) 1815.

8 R.C. Henry, C.W. Lewis, P.K. Hopke and H.J. Williamson, Atmospheric, 18 (1984) 1507.
9 R. Bandeen-Roache and D. Ruppert, Chemom. Intell. Lab. Syst., 10 (1990) 169.
10 K. Bandeen-Roache, in preparation.
11 B. Efron and R. Thisted, Biometrika, 63 (1976) 435.
12 S.-H. Lo, Ann. Stat., 20 (1992) 1094.
13 J. Sacks and D. Ylvisaker, Ann. Stat., 6 (1978) 1122.
14 N. Cressie, C.A. Gotway and M.O. Grondona, Chemom. Intell. Lab. Syst., (1990) 251.
15 D. Donoho, I.M. Johnstone, J.C. Hoch and A.S. Stern, J. Stat. Soc. B., 54 (1992) 41.

Analytica Chimica Acta, 277 (1993) 357–367
Elsevier Science Publishers B.V., Amsterdam

Gas chromatography–pattern recognition techniques in pollution monitoring

B.K. Lavine

Department of Chemistry, Clarkson University, Potsdam, NY 13699–5810 (USA)

A. Stine

Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY 13699–5810 (USA)

H.T. Mayfield

HQ AFCESA / RDVC, Tyndall AFB, FL 32403–6001 (USA)

(Received 28th September 1992)

Abstract

Gas chromatogaphy and pattern recognition methods (GC–PR) constitute a powerful tool for investigating complex environmental problems, e.g., source identification of pollutants. Multivariate methods are necessary to realistically analyze large chromatographic data sets and to seek meaningful relationships between chemical constitution and environmental variables. Recently, our laboratory has investigated the potential of GC–PR as a method for typing jet fuels so a spill sample could be related to its source. Fuels that had undergone weathering in a subsurface environment were correctly identified as to type using disjoint principal component models developed from gas chromatograms of unweathered jet fuels. The constituents of synthetic binary fuel mixtures were also correctly identified by fuel type using the principal component models.

Keywords: Gas chromatography; Pattern recognition; Pollution monitoring; Waters

Water from underground wells or aquifers is an important water resource, supplementing or replacing surface water supplies in many communities in the United States. Possible contamination of this water resource by processed fuels stored in leaking underground tanks or pipelines has prompted the United States Air Force to develop new methods for identifying fuel materials recovered from the subsurface environment on or near military airfields. The Air Force's interest in techniques which can establish the type of fuel responsible for the contamination (and hence the source of the pollutant) is motivated in part by the clean-up costs, legal fees, and fines which could be incurred by the polluter.

Water samples from underground wells contaminated by leaking fuels exist in one of two forms. Either the water sample collected from the well shows layers of floating fuel or the sample contains the dissolved hydrocarbons from the leaking fuel. In the worst case scenario (i.e., that of floating fuel), the fuel layer is collected and analyzed by gas chromatography (GC). The gas chromatographic profile of the fuel layer is then compared with the GC profile of the different candidate fuels to determine the fuel type of the leaking material. In the other case (that of the

Correspondence to: B.K. Lavine, Department of Chemistry, Clarkson University, Potsdam, NY 13699–5810 (USA).

dissolved hydrocarbons), the GC profile of the hydrocarbons is compared to the GC profile of only the water soluble components of a candidate fuel to seek a match.

Typically, the gas chromatograms are compared visually in order to obtain the best match. However, this approach to data analysis is somewhat subjective and is often not persuasive in a court of law. Visual analysis of chromatograms also has the drawback that it cannot always take into account the influence of weathering on the overall GC profile. Weathering [1,2] causes selective evaporation of lower-molecular-weight alkanes and dissolution of aromatics and other water soluble compounds in the fuel. Small variations in the GC operating conditions (e.g., small changes in the temperature programming rate of the GC column) can also be a problem that can complicate the identification of a leaking fuel [3,4].

Pattern recognition methods [5–8] offer a better approach to the problem of matching the gas chromatograms of processed fuels since these methods usually involve less subjectivity in the interpretation of the data. Pattern recognition methods are capable of identifying fingerprint patterns within gas chromatographic data characteristic of a given fuel type even if the fuel samples comprising the training set have been subjected to a variety of conditions. Hence, classifiers can be developed that are relatively insensitive to the changes in the overall GC profile of the original fuel due to contamination, analytical error, or weathering. Furthermore, pattern recognition methods do not make any implicit assumptions about the underlying statistical distribution of the data, and this is particularly advantageous since GC profile data often cannot be approximated by a statistically tractable probability distribution function.

The present work describes the application of pattern recognition methods to the problem of classifying gas chromatograms of weathered and unweathered jet fuels. Our goal is to demonstrate that information about fuel type can also be extracted from the gas chromatograms of weathered fuels. Classifiers developed with gas chromatograms of unweathered jet fuels were used to predict the fuel type of weathered fuels. This study is an extension of an earlier effort [9] which emphasized the development of a nonparametric pattern recognition methodology for the interpretation of GC profile data.

EXPERIMENTAL

Ninety-one fuel samples representing six different types of jet fuels [JP-4, Jet-A, JP-7, JPTS, JP-5, and 100/130 octane aviation gasoline (AVGAS)] were obtained from Wright Patterson Air Force Base and Mukilteo WA Energy Management Lab. (see Table 1). These samples constituted the training set and were splits from regular quality control standards used by these laboratories to verify the authenticity of manufacturers' claims. The quality control standards were collected over a three year period and were a representative sampling of the fuels.

Fifteen fuel samples – each a 50:50 mixture of two different fuels (e.g., JP-4 and JetA) – were prepared from the samples comprising the training set. Since a fuel spill could be the product of two or more fuels, it would be desirable to develop the capability of predicting the class membership values for gas chromatograms of binary fuel mixtures. Hence, these 15 fuels samples were placed in the prediction set (see Table 2).

Eight fuel samples (6 JP-4 and 2 JP-5) that had been recovered from the subsurface environment near Air Force and Navy airfields were also placed in the prediction set (see Table 2). The recovered fuels were collected from sampling wells as a

TABLE 1

Neat jet fuel training set

Number of samples	Fuel-type
34	JP-4 (fuel used by USAF fighters)
15	JetA (fuel used by civilian airliners)
7	JP-7 (fuel used by the USAF SR-71 reconnaissance plane)
7	JPTS (fuel used by USAF TR-1 and U-2 aircraft)
9	JP-5 (fuel used by Navy jets)
19	AVGAS (100/130 octane aviation gasoline)

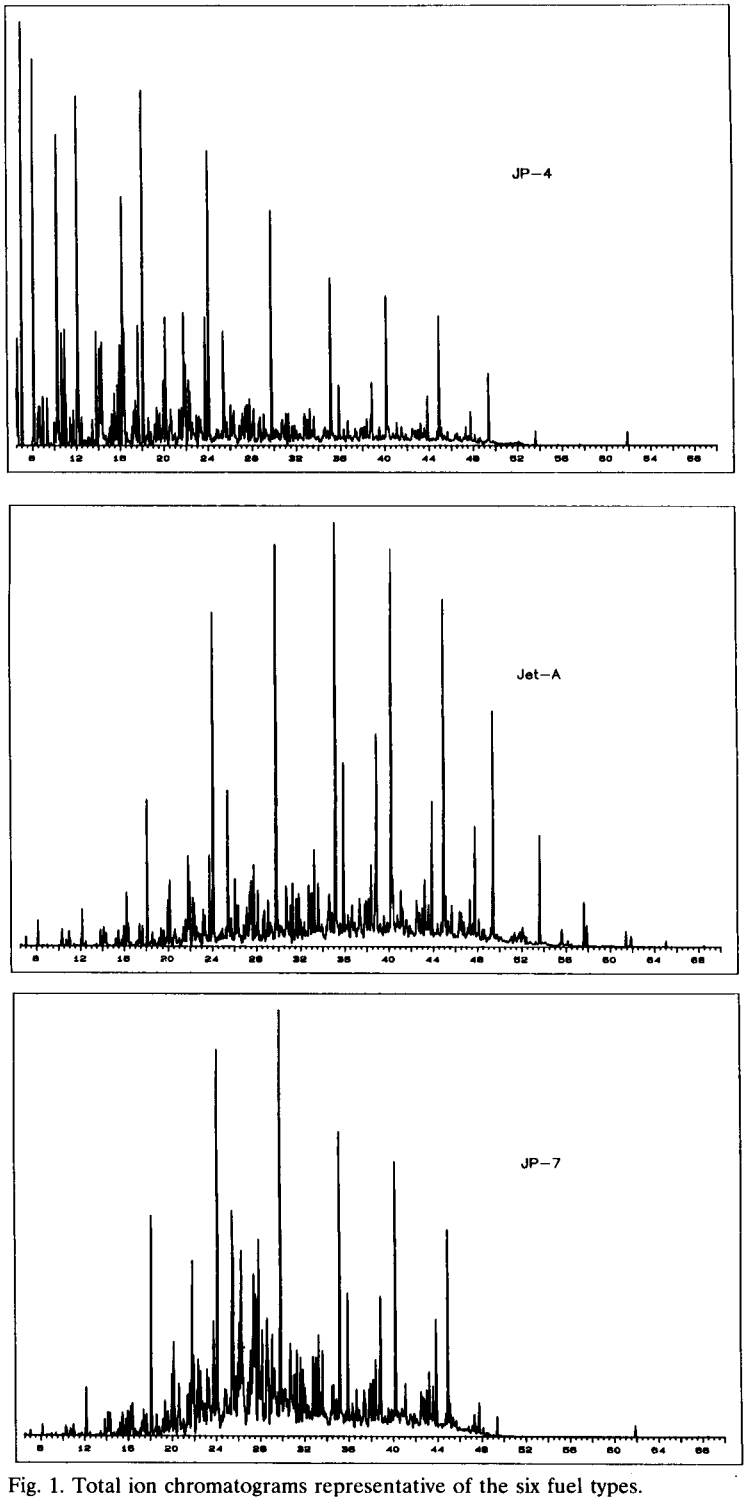


Fig. 1. Total ion chromatograms representative of the six fuel types.

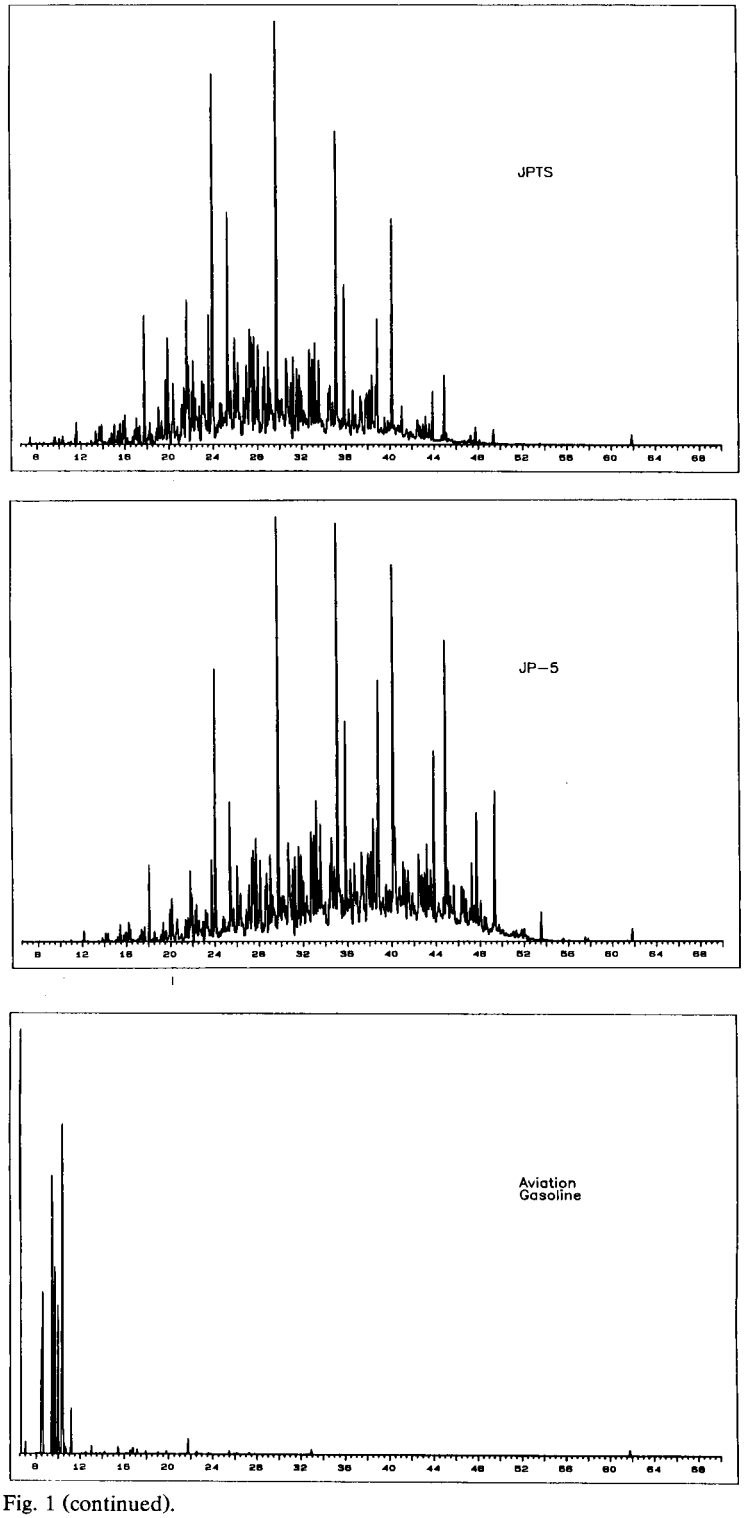


Fig. 1 (continued).

TABLE 2
Prediction set

Sample No.	Description
1	50:50 mixture of JP-4/JetA
2	50:50 mixture of JP-4/JP-5
3	50:50 mixture of JP-4/JP-7
4	50:50 mixture of JP-4/JPTS
5	50:50 mixture of JP-4/AVGAS
6	50:50 mixture of JP-5/JP-7
7	50:50 mixture of JP-5/JPTS
8	50:50 mixture of JetA/JP-5
9	50:50 mixture of AVGAS/JP-5
10	50:50 mixture of JPTS/JP-7
11	50:50 mixture of JetA/JP-7
12	50:50 mixture of AVGAS/JP-7
13	50:50 mixture of JetA/JPTS
14	50:50 mixture of AVGAS/JPTS
15	50:50 mixture of AVGAS/JetA
16	Recovered JP-4 (2 years)
17	Recovered JP-4 (4 years)
18	Recovered JP-4 (4 years)
19	Recovered JP-4 (4 years)
20	Recovered JP-4 (4 years)
21	Recovered JP-4 (4 years)
22	Recovered JP-5 (unknown)
23	Recovered JP-5 (unknown)

neat, oily phase floating on top of the well water. One of the JP-4 fuel samples had been recovered from the subsurface environment two years after the repair of an underground pipeline. The other five JP-4 fuel samples were recovered approximately four years after fuel tanks and hardware had been removed from a petroleum storage area. Since the underground fuel tanks and pipelines may have leaked for a substantial period of time prior to repairs, the figures cited here serve only to fix the minimum age of the weathered samples. There is no available information to estimate the age of the two JP-5 samples, as they were collected shortly after the contamination had been discovered.

Each fuel sample was diluted with methylene chloride prior to gas chromatographic–mass spectrometric (GC–MS) analysis. The diluted fuel sample was spiked with d_{10}-anthracene which served as an internal retention time standard and was then injected onto a GC capillary column (DB-5, J&W Scientific, 30 m × 0.255 mm i.d.) that was temperature programmed from 40 to 250°C at 3°C min^{-1} using an HP 5890 gas chromatograph equipped with an HP-5970 mass selective detector. The mass spectrometer was scanned cyclically over the mass range 35–350 daltons with a 0.66-s scan cycle period. The data were collected and stored on an HP Vector Graphics 386 PC running the Windows operating system and the ChemStation System software. Total ion chromatograms representative of the six fuel types in this study are shown in Fig. 1.

DATA PREPROCESSING AND PATTERN RECOGNITION ANALYSIS

The gas chromatograms were peak matched using the method of Mayfield and Bertsch [10]. First, each gas chromatogram was divided into a number of intervals defined by so-called marker peaks, i.e., peaks that are present in all of the chromatograms. The marker peaks were then peak matched using mass spectral data acquired during the runs. Next, retention time offsets were computed for the marker peaks in each individual chromatogram, i.e., the difference in retention time between the marker peaks in the reference chromatogram (which is selected by the user) and the chromatograms own marker peaks were computed. Finally, all of the peaks between the reference peak markers were adjusted by linear interpolation which is a valid assumption in the case of temperature programming.

To standardize the retention time of peaks eluting prior to the first marker peak, it was necessary to use scaling factors developed for the first pair of marker peaks. Similarly, peaks eluting after the last marker peak used a set of scaling factors developed for the final pair of marker peaks. (These extrapolations are only valid over short time intervals.) Tolerance windows for retention time differences were specified by the user, and the peaks in these intervals were matched with those in the reference chromatogram provided that the difference in the adjusted retention times for a given pair of peaks fell within the specified tolerance window.

The computer program used for peak matching yielded a set of 75 standardized retention

TABLE 3

Retention time windows used for pattern recognition analysis

Feature	Retention time (min)	Kovats Index [a]	Compound identity
1	11.16	785	1,1-Dimethylcyclohexane
2	11.27	787	Unidentified
3	11.81	797	*trans*-1,2-Dimethyl cyclohexane
4	11.94	800	*n*-Octane
5	13.76	831	Propylcyclopentane
6	13.86	833	Unidentified
7	15.15	855	2,3-Dimethylheptane
8	15.44	860	Ethylbenzene
9	15.92	868	*m*- and *p*-xylene
10	16.07	871	3-Ethylheptane
11	16.90	885	Unidentified
12	17.06	888	*cis*-1-Ethyl-3-methyl-cyclohexane
13	17.18	890	Unidentified
14	17.33	893	*o*-Xylene
15	17.76	900	*n*-Nonane
16	18.21	908	Unidentified
17	18.98	921	Unidentified
18	19.24	925	Isopropylbenzene + hydrocarbons
19	19.60	931	Unidentified
20	19.78	934	2,6-Dimethyloctane
21	20.11	940	Unidentified
22	20.23	942	Unidentified
23	20.85	952	Unidentified
24	21.00	955	*n*-Propylbenzene
25	21.18	958	Unidentified
26	21.35	960	Unidentified
27	21.45	962	Unidentified
28	21.59	964	Unidentified
29	21.87	969	1,2,5-Trimethylbenzene
30	21.99	971	3-Methylnonane
31	22.55	981	Unidentified
32	22.78	985	Unidentified
33	22.94	987	Unidentified
34	23.37	994	1,2,4-Trimethylbenzene
35	23.70	1000	*n*-Decane
36	24.04	1006	Unidentified
37	24.21	1009	Isobutylbenzene
38	24.47	1014	Unidentified
39	26.44	1048	Unidentified
40	27.36	1064	2-Methyldecane
41	27.54	1067	1-Methyl-3-propyl-benzene
42	27.72	1071	3-Methyldecane
43	28.16	1078	Unidentified
44	28.45	1083	Unidentified
45	28.62	1086	1,2-Dimethyl-4-ethyl-benzene
46	28.79	1089	Unidentified
47	29.39	1100	*n*-Undecane
48	30.23	1116	Unidentified
49	30.63	1123	1,2,3,4-Tetramethyl-benzene
50	30.86	1128	Unidentified
51	31.16	1133	Unidentified
52	31.39	1138	Unidentified
53	31.59	1141	Unidentified
54	32.30	1155	Unidentified
55	32.57	1160	Unidentified
56	32.79	1164	2-Methylundecane
57	33.15	1171	Unidentified
58	34.71	1200	*n*-Dodecane
59	35.08	1207	Unidentified
60	35.41	1214	Unidentified
61	37.27	1252	Unidentified
62	37.67	1260	Unidentified
63	37.89	1264	Unidentified
64	38.36	1273	Unidentified
65	38.99	1286	Unidentified
66	39.68	1300	*n*-Tridecane
67	41.93	1348	Unidentified
68	42.68	1364	Unidentified
69	43.32	1377	Unidentified
70	44.38	1400	*n*-Tetradecane
71	45.10	1416	2,6-Dimethylnaphthalene + alkanes
72	46.72	1453	2,3-Dimethylnaphthalene + alkanes
73	47.16	1463	Unidentified
74	48.80	1500	*n*-Pentadecane
75	52.97	1600	*n*-Tetradecane
76	61.20	1808	d_{10}-Athracene

[a] Kovats indices were computed for the peaks in each chromatogram using the *n*-alkanes present in the mixtures as retention standards.

time windows (see Table 3). Hence, for pattern recognition analysis, each gas chromatogram was initially represented as a 75 dimensional data vector $\mathbf{X} = (x_1, x_2, x_3, \ldots x_j, \ldots x_n)$ where x_j is the area of the jth peak. The gas chromatograms were normalized to constant sum using total integrated peak area and autoscaled to ensure that each chromatogram and peak had equal weight in the analysis.

In this study principal component analysis (PCA) and other soft modelling methods [11–13]

were used to analyze the data. PCA attempts to reduce the dimensionality of the data by finding a set of orthogonal axes that represent the directions of greatest variance in the data. The orthogonal axes are called the principal components of the data.

PCA is carried out via a decomposion of the data matrix **X** ($n \times p$) into a score matrix **T** ($n \times f$), a loading matrix **P** ($f \times p$), and a residual matrix **E** ($n \times p$). (Note that f is smaller than p.) The matrix equation for the decomposition is

$$\mathbf{X} = \mathbf{1}\mathbf{x}_{mean} + \mathbf{TP} + \mathbf{E} \quad (1)$$

where **1** is a column vector ($n \times 1$) of ones and $\mathbf{x}_{mean}$ is a ($1 \times p$) row vector that is the mean of the observations. The score matrix defines the coordinates of the data points in the principal component space, and the loading matrix defines the relationship between the original measurement variables and the principal components. The score and the loading matrix describe the signal in the data, and the residual matrix describes the noise. Hence, PCA and other soft modelling methods possess two key attributes that will ensure a successful analysis of the data: (1) dimensionality reduction and hence simplification of the multivariate data, and (2) separation of signal from noise in a data matrix.

RESULTS AND DISCUSSION

The first step in this study was to apply PCA to the analysis of the training set data. Figure 2 shows a plot of the scores of the first two principal components of the 75 GC peaks obtained from the 91 neat jet fuel samples. Each sample is represented as a point in the plane defined by the two largest principal components (PC). The two largest principal components explain 73% of the total cumulative variance of the data. The AVGAS, JP-4, Jet-A, JP-7, JPTS, and JP-5 gas chromatograms are reasonably well separated from each other in the PC map, suggesting that information is present in the gas chromatograms that is characteristic of fuel type.

A plot of the loadings, the coefficients for each of the 75 GC peaks used for construction of the principal components, can also reveal information about trends present in the data. Figure 3 shows a plot of the loadings of the first two principal components obtained from the 91 neat jet fuel chromatograms. Each GC peak is represented as a point in the plot, and the location of

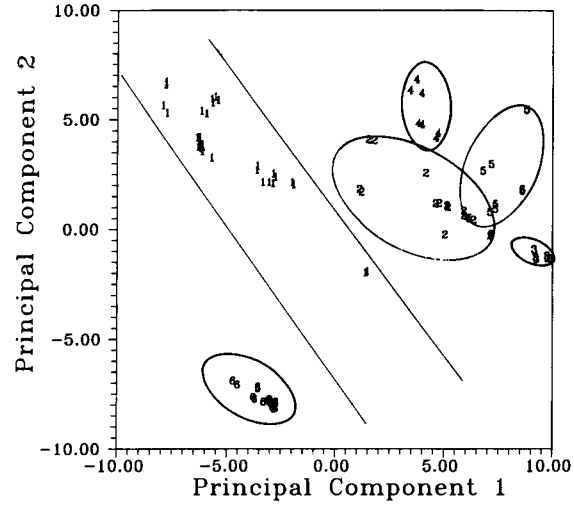


Fig. 2. A principal component map of the training set data. The map was developed from 75 GC peaks and explains 73% of the total cumulative variance. 1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5 and 6 = AVGAS.

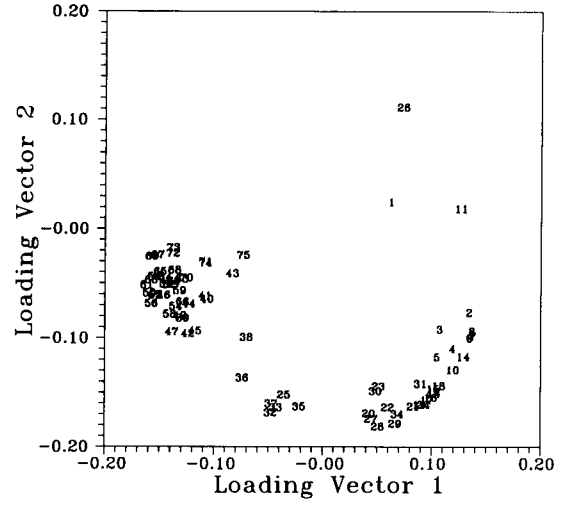


Fig. 3. A plot of the loadings of the two largest principal components of the training set data. 1 = First GC peak in the chromatogram and 75 = the 75th or last GC peak in the chromatogram.

the GC peaks in this plot is consistent with elution order suggesting that confounding of the desired class information by an unwanted experimental artifact, column aging, is a problem. Since the data are highly colinear, it is improbable that one would be able to break apart this confounding effect through an oblique rotation of the data. (Recall that the first two principal components explain 80% of the total cumulative variance of the 75-dimensional data implying a very high condition number for the data matrix.) Of course, if there was a large set of binary and ternary neat jet fuel mixtures of varying composition, e.g., 30% JP-4, 40% Jet-A, 30% JP-7, in the training set, it would be possible to carry out a transformation of the GC data along the mixture line in the pattern space using PLS [14]. Those GC peaks with high loadings on PLS components that define the calibration line would contain little if any information about experimental variables but would contain significant amounts of information about fuel type.

Because there were no gas chromatograms of neat jet fuel mixtures in the training set, the calibration approach to feature selection described above could not be employed. Instead, we used the Fisher (F) ratio [15] to identify GC peaks strongly correlated with fuel type. The F ratio can furnish information about the contribution of a particular GC peak to the solution of a given pattern recognition problem. Therefore, the F ratio was computed for the following category-pairs: JP-4 vs. Jet-A, Jet-A vs. JP-5, Jet-A vs. JP-7, Jet-A vs. JPTS, JP-5 vs. JP-7, and JP-5 vs. JPTS. The total F ratio, which is the arithmetic average of the individual category pair F ratios, was also computed for the set. (We did not include F ratios from AVGAS in the set since this fuel is easily differentiated from the other fuels which is very evident from the principal component plot. Therefore, AVGAS F ratios would not contain unique information about the contribution of a particular GC peak to the pattern recognition problem at hand but would simply inflate the value of the total F ratio.) A GC peak was retained for further study if it had a large total F ratio or if its F ratio for the JP-4 vs. Jet-A classification problem or the Jet-A vs. JP-5 classification problem was very large. (The Air Force was particularly interested in developing the capability of reliably differentiating JP-4 from Jet-A and JP-5 from Jet-A.) Forty-two GC peaks which span the entire gas chromatogram were retained for further study.

The effect of column aging on the classification of the jet fuels in the truncated pattern space was assessed using SIMCA pattern recognition [16]. A PC model for each class in the training set was developed; the complexity of the PC models were determined directly from the data using the technique of cross validation [17]. For each class, a two-component model was developed. The gas chromatograms in the training set were fitted to these models, and the residual, which is the sum of the squares of the difference between the original chromatogram and the gas chromatogram reproduced by the PC model, was computed.

Each gas chromatogram in the training set was then classified on the basis of its residual or goodness of its fit. In other words, the probability of a gas chromatogram belonging to a particular class of fuels was determined from the residual for the corresponding PC model by way of an F

TABLE 4

SIMCA training set results

Class	Number of PCs	Number in class	Right [a]	Wrong [b]	Right (%)
JP-4	2	34	34	0	100.00
JetA	2	15	14	1	93.33
JP-7	2	7	7	0	100.00
JPTS	2	7	7	0	100.00
JP-5	2	9	9	0	100.00
AVGAS	2	19	19	0	100.00
Total		91	90	1	98.90

[a] Classifications were made on the basis of the variance ratio: $F = [s_p / S_o]^2 [N_q / (N_q - NC - 1)]$, where s_p^2 is the residual of sample p for class i, S_o^2 is the variance within class i, N_q is the number of samples in the class, and NC is the number of principal components used to model the class. A sample was assigned to the class for which it had the lowest variance ratio. However, if the variance ratio exceeded the critical F value for that class, then the sample was not assigned to it. [b] 90 of the 91 samples in the training set were uniquely classified (see Ref. 16 for definition of uniquely classified).

TABLE 5

SIMCA prediction set results

Sample No.	Fuel type	Predicted class membership
1	JP-4/JetA	JP-4
2	JP-4/JP-5	– [a]
3	JP-4/JP-7	JP-4/JP-5 [b]
4	JP-4/JPTS	–
5	JP-4/AVGAS	JP-4
6	JP-5/JP-7	JP-5
7	JP-5/JPTS	JP-5
8	JetA/JP-5	JP-5
9	AVGAS/JP-5	–
10	JPTS/JP-7	JP-5
11	JetA/JP-7	JetA/JP-5
12	AVGAS/JP-7	JetA/JP-5
13	JetA/JPTS	JP-4/JPTS
14	AVGAS/JPTS	JP-5
15	AVGAS/JetA	JP-4
16	JP-4 (2 years)	JP-4
17	JP-4 (4 years)	JP-4
18	JP-4 (4 years)	JP-4
19	JP-4 (4 years)	JP-4
20	JP-4 (4 years)	JP-5
21	JP-4 (4 years)	Jet-A
22	JP-5	JP-5
23	JP-5	JP-5

[a] Since the variance ratio exceeded the critical F value for each fuel category at $\alpha = 0.01$, the sample was not assigned to any of the classes, i.e., it was labeled a non-member. Critical F values for prediction set samples were obtained using one degree of freedom in the numerator and $N_q - NC - 1$ degrees of freedom in the denominator (see Ref. 21). [b] If a prediction set sample has a variance ratio less than the critical value for two fuel classes, the sample in question possesses dual class membership. A similar statement would hold true for three classes, etc.

test. In addition, the PC models were validated using an external prediction set of 23 gas chromatograms (see Table 2). The chromatograms constituting this external prediction set were run a year after the neat jet fuel chromatograms were run and thus constituted a true prediction set. The results of the SIMCA pattern recognition study are summarized in Tables 4 and 5. The classification success-rate for the neat jet fuels was 98.9% (90:91), and 75% for the weathered fuels. (6 out of 8 weathered jet fuel chromatograms were correctly classified.) However, the classification success-rate for the mixtures was low.

The failure to classify the mixtures was a concern because floating fuel found in a well can be the product of two or more fuels. Hence, the FCV clustering algorithm [18] was also used in this study to analyze the data. The FCV clustering algorithm attempts to fit each of the c-classes in the data to a principal component model. One interesting feature of the FCV algorithm is that each data vector in the training set contributes to the modelling of each of the classes within the data. The FCV clustering algorithm is, therefore, better equipped to tackle the problem of dual class membership than traditional pattern recognition methods, e.g., SIMCA.

The FCV clustering algorithm consists of four equations that must be solved simultaneously by way of a Picard iteration [19].

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}} \right)^{1/(m-1)}} \tag{2}$$

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} (\mu_{ik})^m} \tag{3}$$

$$S_i = \sum_{k=1}^{n} (\mu_{ij})^m (x_k - v_i)(x_k - v_i)^T \tag{4}$$

$$D_{ik} = \left(|x_k - v_i|^2 - \sum_{j=1}^{r} \langle x_k - v_{ij}, d_{ij} \rangle^2 \right)^{1/2} \tag{5}$$

μ_{ik} is the membership value of sample k with respect to class i ($i = 1,2,3,\ldots c$), and it is subject to the following boundary conditions: $0 < \mu_{ik} < 1$ and $\Sigma\mu_{ik} = 1$. D_{ik} is the distance between sample k and cluster center i, v_i is the center of cluster i, d_{ij} is a unit eigenvector corresponding to the jth largest eigenvalue of the fuzzy covariance matrix S_i, m is a fixed weighting exponent (usually set at 2), and r defines the shape of the point cluster ($r = 0$ for spherical, $r = 1$ for linear varieties and so forth).

Cluster analysis was performed with the samples in the training set using the 42 autoscaled variables. The centroids of each fuel class were

the starting centers for the analysis. r was set equal to 0, m was set equal to 2, and c was varied from 6 to 10. The cluster validity coefficient (CVC) which is a measure of the separation between clusters was also computed for each cluster configuration investigated. The CVC [20] can be interpreted as the distance between the centers of adjacent clusters relative to the scatter of the two clusters about their respective centers. When $c = 8$, an interesting result was obtained. Clustering on the basis of fuel type was observed (see Table 6). In addition, this particular cluster configuration ($c = 8$) yielded a high cluster validity coefficient (CVC) for every cluster pair suggesting that the eight clusters were well separated.

The principal component models developed in this experiment ($m = 2$, $r = 0$, and $c = 8$) were validated using the gas chromatograms from the prediction set. The chromatograms in the prediction set were fitted to the linear models, and class membership values for each fuel sample were computed. A fuel sample was assigned to the cluster for which it had a large class membership value (large being defined as 0.20 or greater for the prediction set samples). The class membership value of a sample that is completely fuzzy, i.e., a sample with equal class membership in all 8 clusters, is 0.125, and 0.20 is 50% larger than this baseline value which is how the threshold level was computed a priori.

If a sample had a large class membership for two clusters, then the sample in question possessed dual class membership. On the other hand, it is possible for a sample not to be assigned to any of the clusters because its class membership value did not exceed the specified threshold level.

TABLE 6

FCV training set results

Number of samples	Fuel type	Maximum membership cluster
9	JP-4	1
3	JetA	
23	JP-4	2
12	JetA	3
2	JP-4	
7	JP-7	4
7	JPTS	5
5	JP-5	6
4	JP-5	7
19	AVGAS	8

TABLE 7

FCV prediction set results

Sample No.	Fuel type	Predicted class membership
1	JP-4/JetA	JP-4/JetA
2	JP-4/JP-5	JP-4
3	JP-4/JP-7	JetA
4	JP-4/JPTS	JP-4/JPTS
5	JP-4/AVGAS	JP-4/AVGAS
6	JP-5/JP-7	JP-5/JP-7
7	JP-5/JPTS	JP-5/JPTS
8	JP-5/JetA	JP-5
9	AVGAS/JP-5	AVGAS/JP-5
10	JPTS/JP-7	JP-5/JetA
11	JetA/JP-7	JetA/JP-5
12	AVGAS/JP-7	JetA/JP-5
13	JetA/JPTS	JetA/JP-5
14	AVGAS/JPTS	JPTS
15	AVGAS/JetA	JetA
16	JP-4 (2 years)	JP-4
17	JP-4 (4 years)	JetA
18	JP-4 (4 years)	JP-4
19	JP-4 (4 years)	JP-4
20	JP-4 (4 years)	JetA/JP-5
21	JP-4 (4 years)	JP-4
22	JP-5	JP-5
23	JP-5	JP-5

The prediction set results are summarized in Table 7. Six of the eight weathered fuel samples were correctly classified. Since the weathered fuel samples used in this study were recovered from the subsurface environment, the predominant factor influencing the weathering process is not evaporation of lower-molecular-weight alkanes. Instead, it is biodegradation due to the action of microbial organisms which evidently does not have too marked an influence on the overall GC profile of the fuels.

Six of the fifteen binary fuel mixtures were correctly classified, i.e., the PC models could identify both fuel components present in each of these six samples. Six of the binary fuel mixtures were only partially classified, that is the principal component models could correctly identify only one of the two fuel components present in each

of these mixtures. However, the principal component models could not identify any of the components present in three of the mixtures. The three mixtures in question all contained JP-7 fuel, and it is the only fuel with very low aromatic content. Since the specifications for JP-7 (e.g., flash point, distillation curve, freezing point, and vapor pressure) are very different from the other fuels, perhaps there was a problem with preparing binary fuel mixtures that contained JP-7.

Conclusions

Fuels that have undergone weathering in a subsurface environment have been classified correctly as to type by disjoint principal component models developed from GC data of neat jet fuels. The components of synthetic binary fuel mixtures could also be correctly classified by fuel type using the PC models. Although the confounding of chemical information with experimental variables is a source of difficulty, principal component models have been developed that we believe classified the samples primarily on the basis of valid chemical differences between the different classes of fuels.

The author thanks Janet M. Lavine for proofing the manuscript. This study was supported by Contract Number F08635-90-C-0105 between Clarkson University and the United States Air Force.

REFERENCES

1 Z.R. Regnier and B.F. Scott, Environ. Sci. Technol., 9 (1975) 469.
2 W.C. Yang and H. Wang, Water Res., 11 (1977) 879.
3 B.K. Lavine, X. Qin, A. Stine and H.T. Mayfield, Proc. Control Quality, 2 (1992) 347.
4 J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine and A.M. Harper, Anal. Chem., 1 (1985) 295.
5 J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, Reading, MA, 1974.
6 P.C. Jurs and T.L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
7 K. Varmuza, Pattern Recognition in Chemistry, Springer-Verlag, Berlin, Heidelberg, New York, 1980.
8 D. Coomans and I. Broeckaert, Potential Pattern Recognition in Chemical and Medical Decision Making, Research Studies Press, Wiley, Letchworth, 1986.
9 B.K. Lavine, Chemom. Intell. Lab. Syst., 15 (1992) 219.
10 H.T. Mayfield and W. Bertsch, J. Comput. Appl. Lab., 1 (1983) 130.
11 I.T. Jolliffe, Principal Components Analysis, Springer-Verlag, New York, 1986.
12 J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
13 S. Wold, C. Albano, W.J. Dunn, K. Esbensen, K. Hellberg, E. Johanson and M. Sjostrom, in H. Martens and H. Russwurm (Eds.), Pattern Recognition: Finding and Using Regularities in Multivariate Data, in Food Research and Data Analysis, Applied Sciences, London, 1983, pp. 147–188.
14 H. Martens and T. Naes, Multivariate Calibration, Wiley, New York, 1989.
15 R.A. Fisher, Ann. Eugen., 7 (1936) 178.
16 S. Wold and M. Sjostrom, in B. Kowalski (Ed.), Chemometrics: Theory and Practice, ACS Symposium Series, No. 52, American Chemical Society, Washington, DC, 1977, pp. 243–282.
17 S. Wold, Technometrics, 20 (1978) 397.
18 J.C. Bezdek, C. Coray, R. Gunderson and J. Watson, SIAM J. Appl. Math., 40 (1981) 358.
19 E. Kreyszig, Advanced Engineering Mathematics, Wiley, New York, 4th edn., 1979, pp. 54.
20 R.W. Gunderson and K.E. Thrane, in J.J. Breen and P.E. Robinson (Eds.), Environmental Applications of Chemometrics, ACS Symposium Series, No. 292, American Chemical Society, Washington, DC, 1985, pp. 130–148.
21 P.J. Gemperline, L.D. Weber and F.O. Cox, Anal. Chem., 61 (1989) 138.

Analytica Chimica Acta, 277 (1993) 369–380
Elsevier Science Publishers B.V., Amsterdam

Potential source contribution function analysis and source apportionment of sulfur species measured at Rubidoux, CA during the Southern California Air Quality Study, 1987

Ning Gao, Meng-Dawn Cheng and Philip K. Hopke
Department of Chemistry, Clarkson University, Potsdam, NY 13699 (USA)

(Received 3rd September 1992)

Abstract

During the Southern California Air Quality Study (SCAQS) in summer and fall of 1987, specially designed SCAQS samplers were used to collect particles and gaseous species. Ion chromatography (IC) and colorimetry were employed to analyze the gaseous and particulate ionic species while the trace elements in the particles were analyzed using x-ray fluorescence (XRF). Potential source contribution function (PSCF) analysis was applied to the chemistry data of the acidic species collected by the SCAQS samplers at the Burbank, Claremont and Rubidoux sites and the meteorology data in the form of air parcel backward trajectories. The results are presented as gridded conditional probability maps showing source areas that have a potential to contribute to the high concentrations of acidic species observed at the receptor sites. In order to quantitatively determine the mass contributions of acidic species from these identified source areas to the receptor sites, a source apportionment method was developed that utilizes the PSCF analysis results and the ground and elevated level emission inventories. One result is a gridded joint probability map showing the quantities of emitted acidic species from source areas that were transported to the three receptor sites with and/or without undergoing chemical transformations during the transport. Another result is a gridded map for each single receptor site showing the possible maximum amounts of emitted acidic species from source areas that were transported to that receptor site with and/or without undergoing chemical transformations. The potential source areas identified by the PSCF analysis and the amount of transported emissions identified by the source apportionment method are generally well correlated with the emission inventories. The receptor modeling results of SO_2 and SO_4^{2-} are presented to illustrate the methodology.

Keywords: Ion chromatography; X-ray fluorescence spectrometry; Potential source contribution function analysis; Southern California Air Quality Study; Sulphur

The objectives of this work are to investigate the sources, transportation, and fate of acidic atmospheric components in the South Coast Air Basin (SoCAB) as measured during the field experiment stage of the Southern California Air Quality Study (SCAQS) in summer and fall of 1987. To examine the source–receptor relationships in the basin, potential source contribution function (PSCF) analysis has been applied to all of the acidic species concentration data for gaseous and particulate samples collected at the Burbank, Claremont, and Rubidoux sites. These sites are shown in Fig. 1 along with a contour plot

Correspondence to: P.K. Hopke, Department of Chemistry, Clarkson University, Potsdam, NY 13699 (USA).

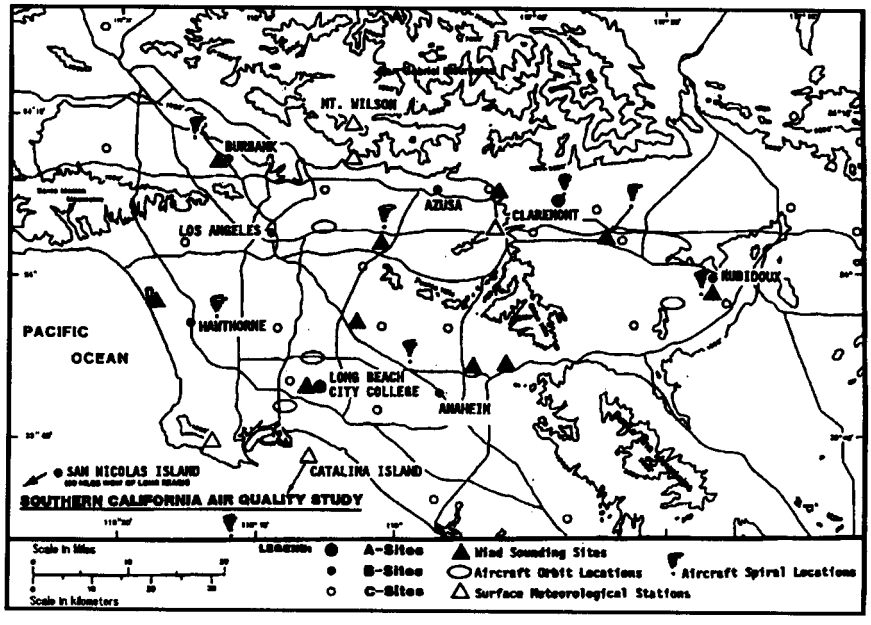


Fig. 1. Map of the Southern California Air Quality Study Region.

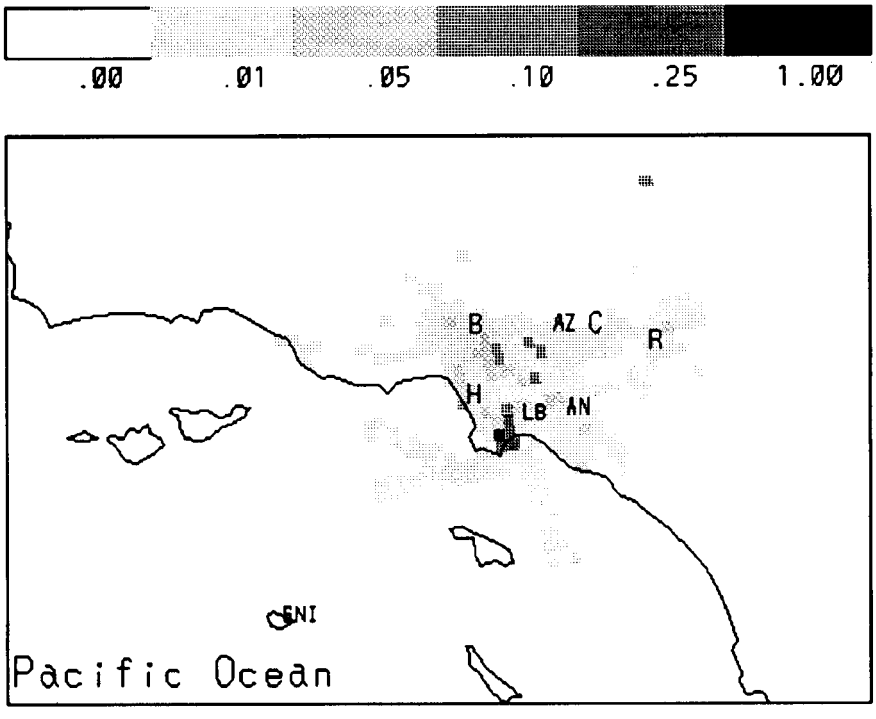


Fig. 2. SO_2 ground level source hourly averaged emission (ton h^{-1}), August 27, 1987.

of the surrounding terrain. The source apportionment method has been subsequently applied for some of these acidic species. The data analysis methods provide an improved understanding of the nature and behavior of the acidic species observed in the SoCAB during the SCAQS field exercises. In this report, the results for SO_2 and SO_4^{2-} observed at Rubidoux will be presented to illustrate the analytical methodology.

DESCRIPTION OF DATA

The SCAQS samplers were designed and built for SCAQS and are discussed by Wolff et al. [1]. The SCAQS sampler consists of 12 sampling lines. Line 1 contains a PTFE–nylon filter sandwich that is analyzed for total nitrate (total NO_3^-). Lines 2 to 9 follow a PM2.5 inlet. Line 2 contains a PTFE filter followed by a carbonate-impregnated cellulose filter. The PTFE filter removes particulate matter, while the cellulose filter is analyzed for SO_2 and SO_4^{2-}. Line 3 contains a denuder coated with magnesium oxide that removes HNO_3. It is followed by a PTFE–nylon filter sandwich that is analyzed for fine particulate NO_3^- (NO_3^- PM2.5 L3). Line 4 contains a PTFE–nylon filter sandwich that yields the sum of gaseous HNO_3 and fine particulate NO_3^-. The concentration of gaseous HNO_3 is obtained by subtracting line 3 from line 4. Line 5 contains a denuder coated with oxalic acid that removes gaseous NH_3 and an oxalic acid impregnated filter that is analyzed for fine particulate NH_4^+ (NH_4^+ PM2.5 L5). Line 6 was not used. Line 7 contains a pre-fired Pallflex quartz filter that is analyzed for organic carbon (OC) and fine elemental carbon (EC). Line 8 contains a tared PTFE filter and is used to determine fine particle mass (Mass PM2.5). Line 9 contains a PTFE filter followed by a quartz filter. The PTFE filter is extracted and analyzed for ionic species (NH_4^+ PM2.5 L9, Cl^- PM2.5, SO_4^{2-} PM2.5, and NO_3^- PM2.5 L9). The quartz filter is analyzed for artifact organic carbon (AO) that results from the absorption of organic vapors. Sampling lines 10 to

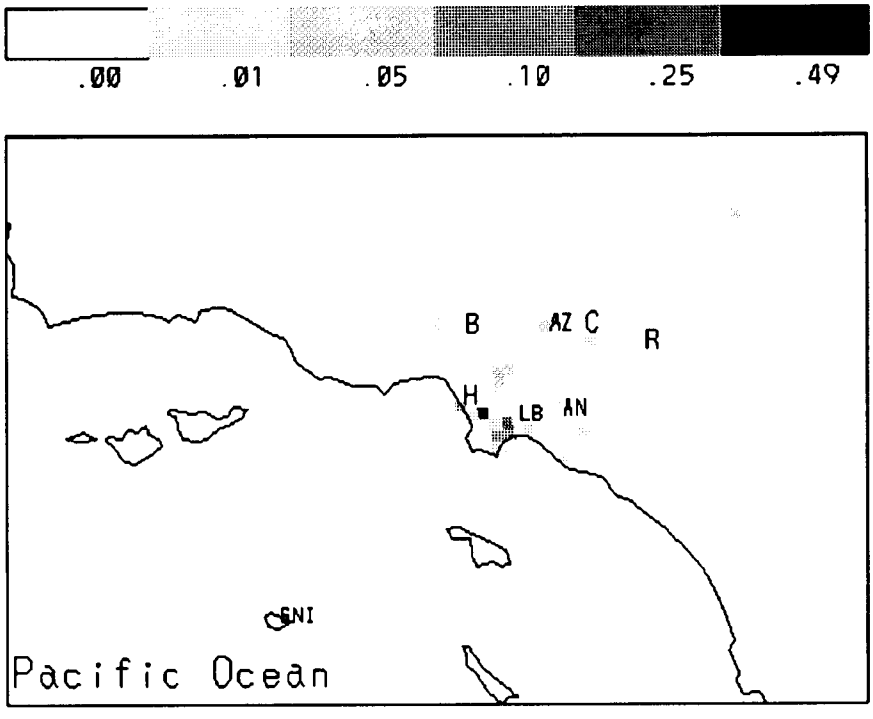


Fig. 3. SO_2 elevated level source hourly averaged emission (ton h^{-1}), August 27, 1987.

12 follow a PM10 inlet. EC PM10 and the sum of OC PM10 and AO are obtained from line 10 that is equipped with a quartz filter. PM10 mass is obtained form line 11 that contains a PTFE filter. The PM-10 concentrations of the ions (NH_4^+ PM10, SO_4^{2-} PM10, NO_3^- PM10, and Cl^- PM10) are obtained from the PTFE filter in line 12.

The fine and PM-10 anions (SO_4^{2-}, NO_3^-, Cl^-, and SO_2 analyzed as SO_4^{2-}) collected with the SCAQS samplers were determined using ion chromatography. The fine and PM-10 ammonium species (NH_4^+ and NH_3 analyzed as NH_4^+) were analyzed using colorimetry. The PM-10 sodium was determined by atomic absorption. The PM2.5 and PM10 masses were measured gravimetrically. These species include gaseous SO_2, SO_4^{2-} (PM2.5, PM10), gaseous HNO_3, NH_3, and NO_3^- (= gaseous $HNO_3^- + PM2.5\ NO_3^-$), NO_3^- (PM10, PM2.5 L3 and PM2.5 L9), NO_3^- (total), NH_4^+ (PM10, PM2.5 L5 and PM2.5 L9), Cl (PM10 and PM2.5), PM10 Na, and mass (PM2.5, PM10). Eighty-five samples each were collected at Burbank and Rubidoux on the 17 intensive study days during the summer and fall of 1987, and 55 samples were collected at the Claremont site during the 11 intensive study days in summer of 1987.

For the IC, AA and colorimetry samples, the below detection limit data (concentration $< 0\ \mu g\ m^{-3}$) were replaced with $0\ \mu g\ m^{-3}$. Missing values were replaced with the mean concentrations of the respective species.

DATA ANALYSIS

Potential source contribution function analysis

To trace the acidic species to their emission source areas, air parcel movement in the form of calculated back trajectories has been combined

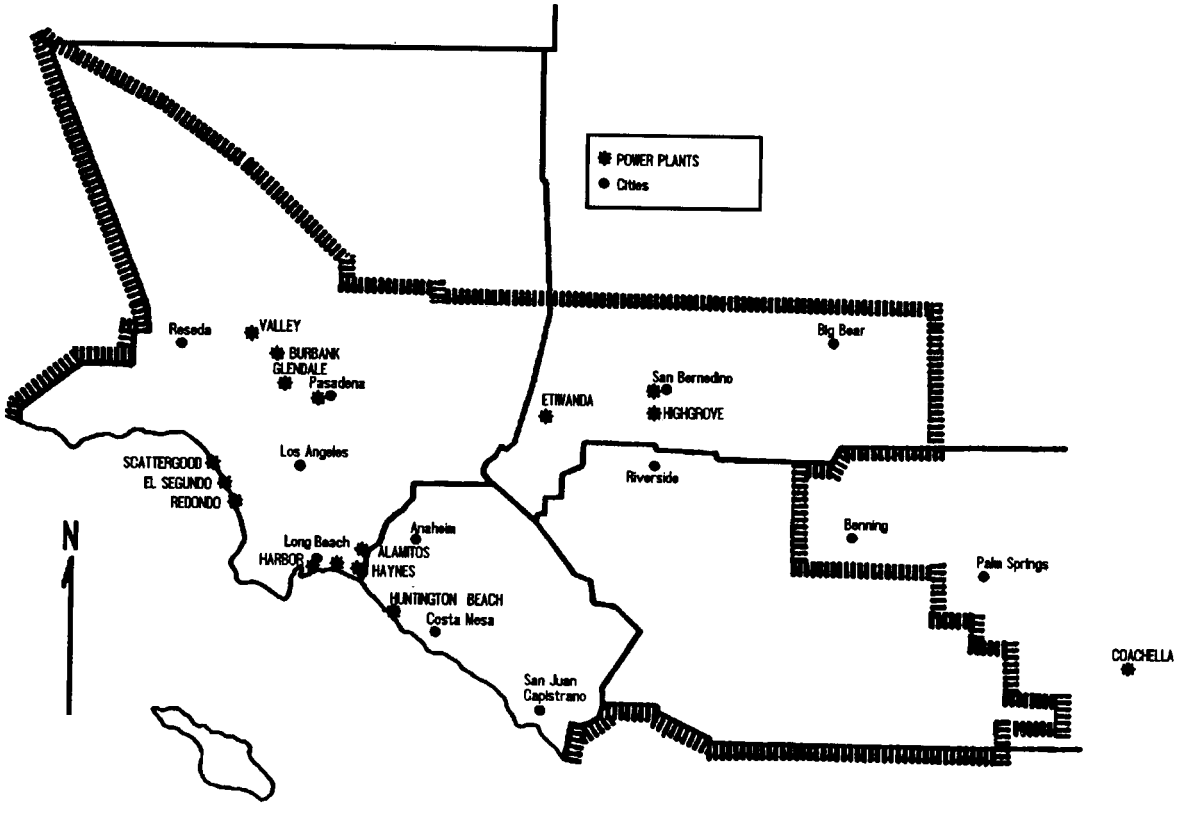


Fig. 4. Map showing the locations of the electricity generating stations in and around the South Coast Air Basin [6].

with the measured chemical species using potential source contribution function analysis [2,3]. The potential source contribution function (PSCF) is the probability that an air parcel with a given level of pollutant concentration or higher arrives at a receptor site after having been passed through a specific geographical area. Air parcel backward trajectories ending at each of the three receptor sites were calculated by Aero Vironment Inc., using a model developed at the California Institute of Technology [4]. The model uses a terrain-following path at 100 m above ground. The trajectories were computed for each hour backward in time over a 24-h period. Thus, there are 24 1-h trajectories for each of the 17 intensive study days. Each 4, 5, or 7 h sample thus has 96, 120, or 168 segment endpoints, respectively, associated with it. The region is subdivided into an i by j gridded array. Let N represent the total number of trajectory segment endpoints during the study period. If n_{ij} endpoints fall into the ijth cell, the probability of an event (i.e., a sample taken over a finite time interval), A_{ij}, is given by

$$P[A_{ij}] = \frac{n_{ij}}{N} \tag{1}$$

In the same ijth cell, if there are m_{ij} endpoints that correspond to the trajectories that arrived at a receptor site with pollutant concentrations higher than some specified criterion value, then the probability of this event, B_{ij}, is

$$P[B_{ij}] = \frac{m_{ij}}{N} \tag{2}$$

The potential source contribution function is then defined as a conditional probability:

$$\mathrm{PSCF}_{ij} = P_{ij} = \frac{P[B_{ij}]}{P[A_{ij}]} = WT\frac{m_{ij}}{n_{ij}} \tag{3}$$

where WT are arbitrary chosen weighting values for cases where n_{ij} is small. In this study, a weight of 0.5 is given to PSCF for the case of

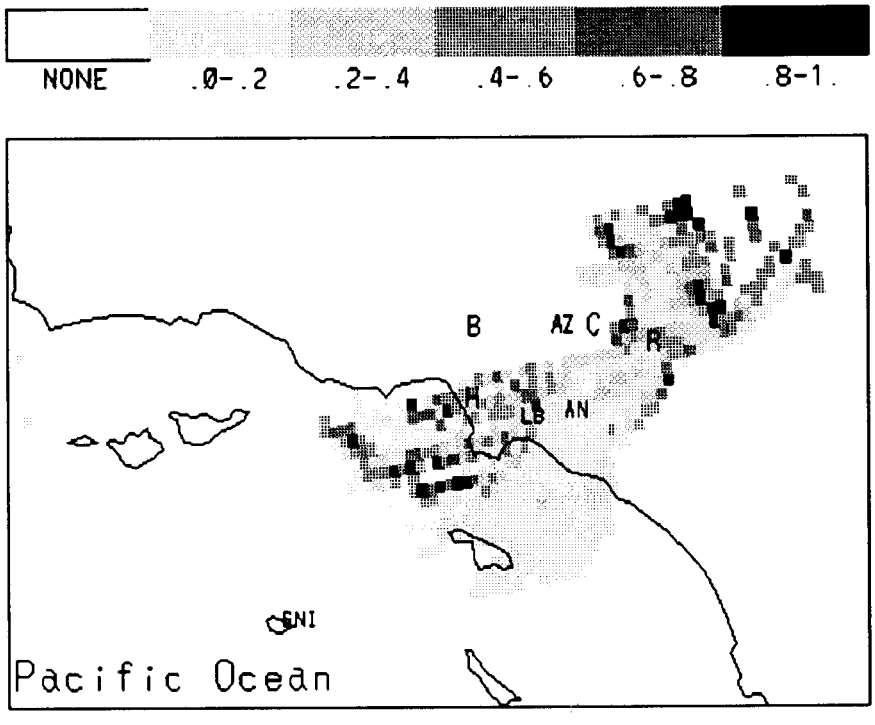


Fig. 5. PSCF map of gaseous SO_2 at Rubidoux.

$n_{ij} = 1$, a weight of 0.68 for $n_{ij} = 2$, a weight of 0.85 for $n_{ij} = 3$, and a weight of 1.00 for $n_{ij} \geq 4$. Grid cells containing pollutant sources are expected to have high conditional probabilities. Therefore, the conditional probability function, PSCF, will identify those source areas that have a potential to contribute to the high concentrations of contaminants observed at the receptor site.

The PSCF analyses were performed for each of the sixteen chemical species analyzed in the samples collected at the Burbank, Claremont, and Rubidoux sites plus one computed species, non-marine SO_4^{2-}. Non-marine sulfate in the PM10 size range was calculated by subtracting 0.25 PM10 Na value from SO_4^{2-} for each sample where 0.25 is the mass concentration ratio of SO_4^{2-} to Na in sea water [5]. Figures 2 and 3 show the 1987 gridded emission inventories (in ton h^{-1}) for ground and elevated level SO_2 sources, respectively. The emission inventory maps were produced using the data supplied by Professor A.G. Russell of Carnegie-Mellon University. Figure 4 shows the locations of the major electricity generating stations in the SoCAB region [6].

The PSCF values were computed and displayed using the same 5 km by 5 km grid size used in the trajectory calculations and the mapping of the 1987 emission source inventories. The 75th percentile values in the individual species distributions were chosen as the criterion for the variables examined in the PSCF analyses.

The PSCF results for SO_2 are shown in Fig. 5 while the SO_4^{2-} results are presented in Figs. 6 and 7. The contribution of SO_2 from off-shore areas is visible in the PSCF map. These areas are identified in the emissions inventories as emissions from oil-fired boilers on large ships. The SO_2 could also be produced from the oxidation of dimethylsulfide (DMS) emitted from the ocean. The DMS may be the product of the digestion of waste sludge in the surface sediments. Some of the off-shore areas are ones in which sewage sludge has been discharged for many years [7]. The SO_2 can be further oxidized to SO_4^{2-}. The involvement of the off-shore ocean dumping with

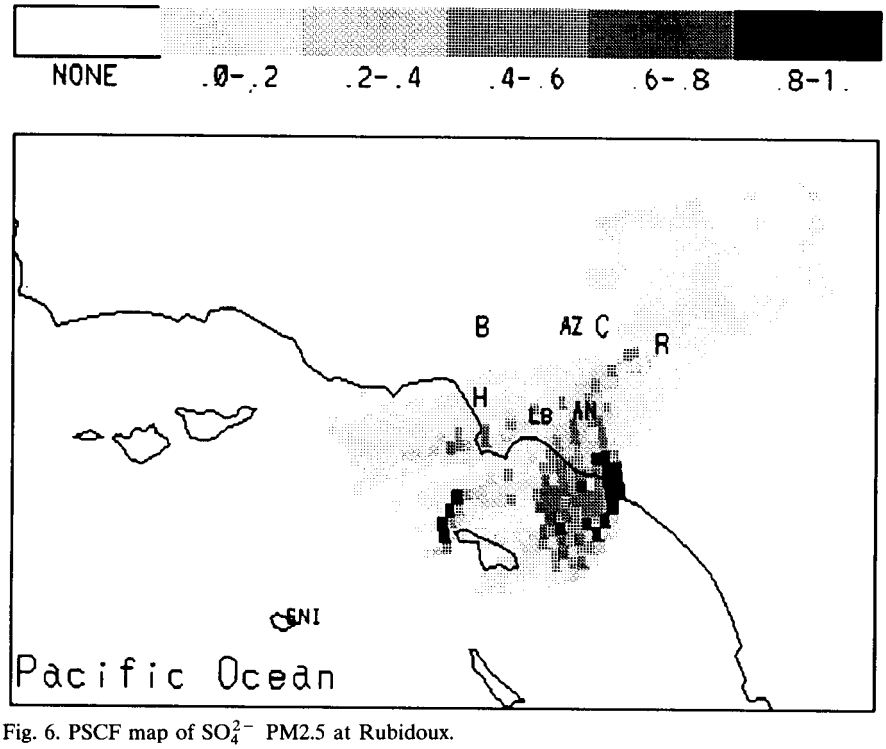


Fig. 6. PSCF map of SO_4^{2-} PM2.5 at Rubidoux.

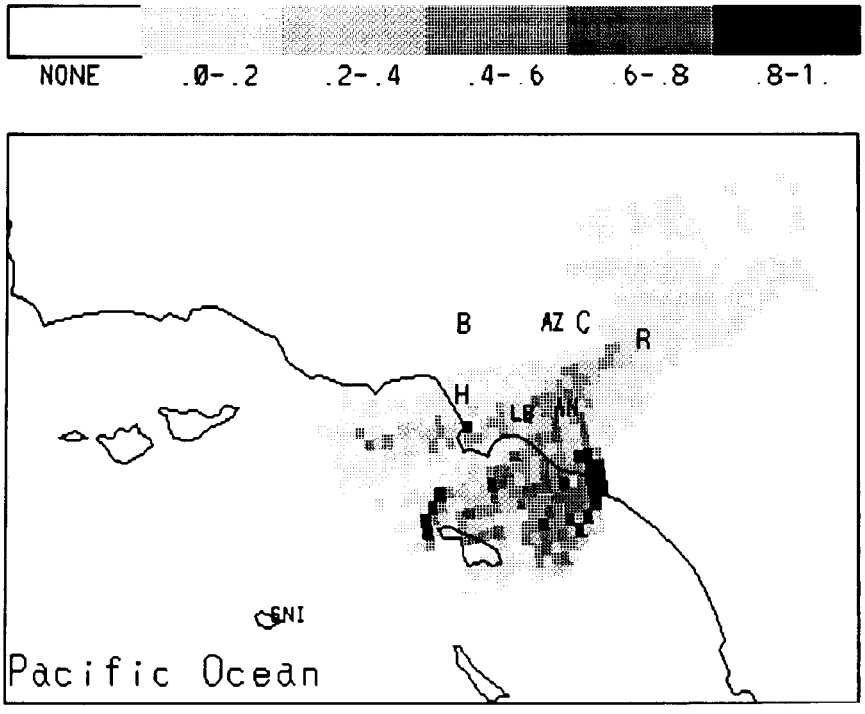


Fig. 7. PSCF map of SO_4^{2-} PM10 at Rubidoux.

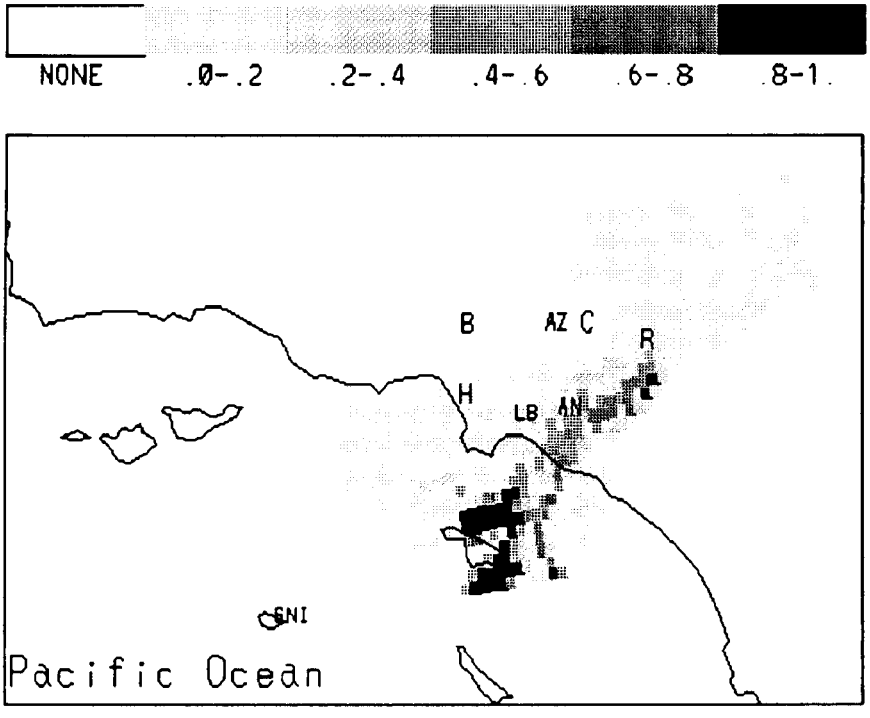


Fig. 8. PSCF map of Na PM10 at Rubidoux.

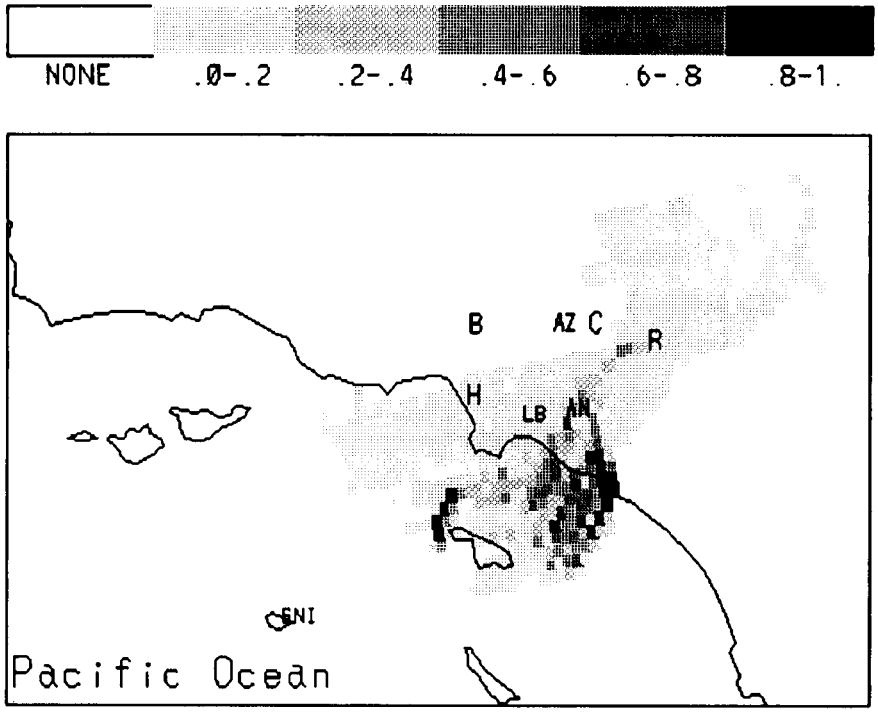


Fig. 9. PSCF map (average concentration as the cut-off value) of non-marine SO_4^{2-} at Rubidoux.

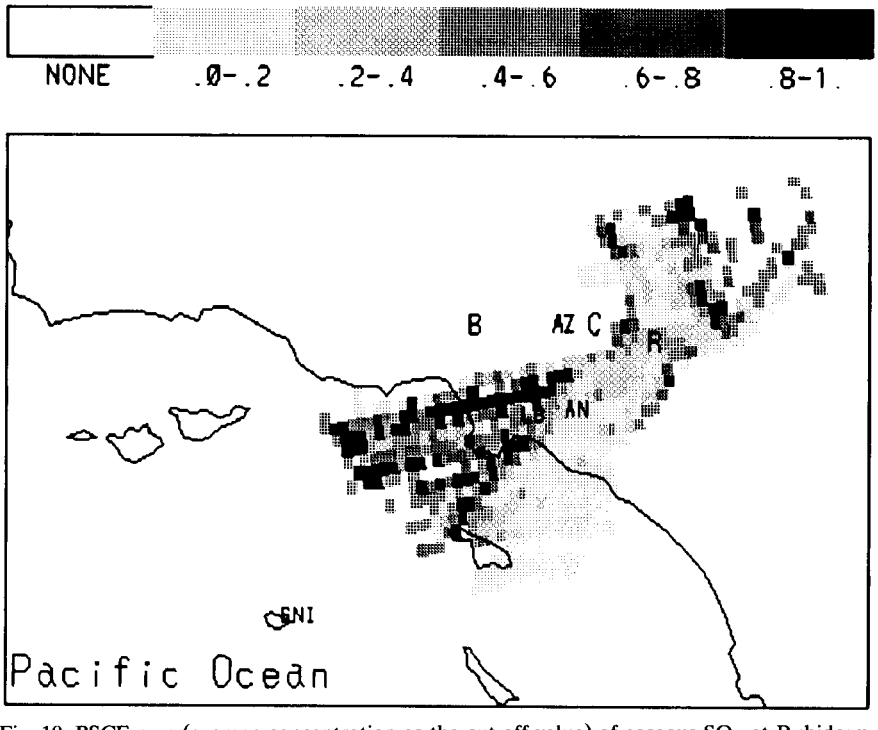


Fig. 10. PSCF map (average concentration as the cut-off value) of gaseous SO_2 at Rubidoux.

production of sulfur species is suggested because some medium and high potential grid cells in the PM10 SO_4^{2-} PSCF map (Fig. 7) does not correlate with those in the PM10 Na PSCF map (Fig. 8). These medium and high potential grid cells are located in the area where the two off-shore waste disposal systems of Los Angeles and Orange county are [7]. The non-marine SO_4^{2-} PSCF map (using average concentration as the cut-off value) also reveals the significant contribution from the Orange county off-shore waste disposal area (Fig. 9). These results suggest that some of the observed SO_4^{2-} found at Rubidoux whose source appears to be some off-shore areas may not be primary sea salt sulfate. This portion of sulfate may be from the oxidation of DMS emitted by the surficial sediments as well as from oxidation of SO_2 emitted by ships. The San Bernedino power plant emits SO_2 at the rate of > 0.01 ton h^{-1} and is located north of Rubidoux. Because of its close proximity to the sampling site, this power plant does not contribute particulate SO_4^{2-} to the sampling site. The widely spread high potential grid cells shown in the SO_2 PSCF map are thought to be caused by the northern trajectory paths leading to Rubidoux that exist during the fall sampling period. The contributions of SO_2 from the power plants and refineries along the areas of Redondo, El Segundo, Scattergood, and Long Beach were masked by the high cut-off value, the 75 percentile, in the PSCF calculation. A new SO_2 PSCF map was thus produced using average concentration as the cut-off value in the calculation. The contributions of these power plants and refineries are revealed more clearly (Fig. 10).

There is reasonable qualitative agreement between the identified regions in the PSCF maps and the known emission locations. However, this analysis does not permit the estimation of quantitative source–receptor relationships and so additional analysis is needed.

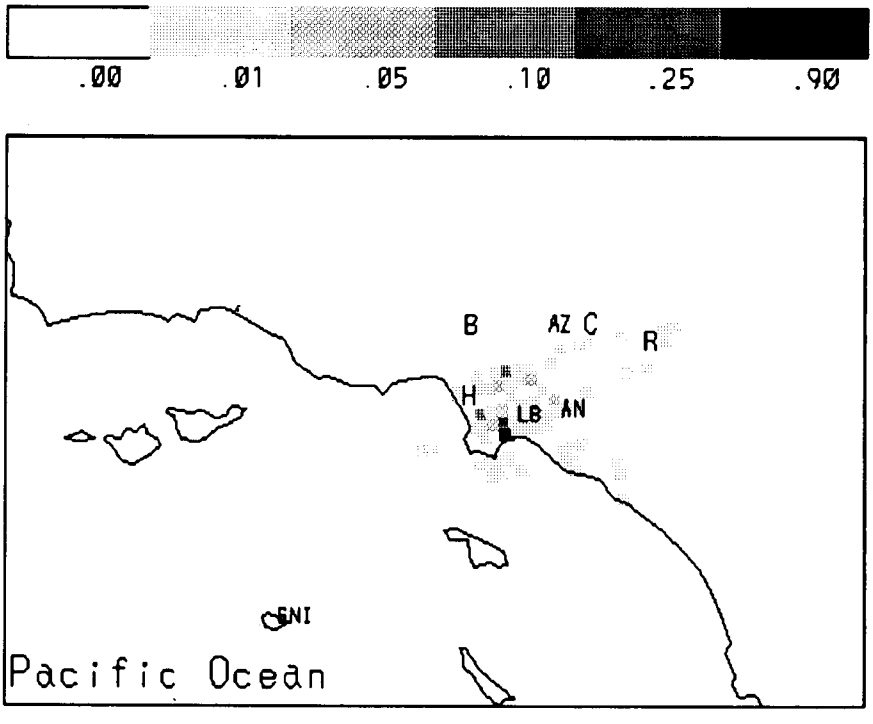


Fig. 11. Source apportionment map of gaseous SO_2 for Rubidoux.

Source apportionment derived from potential source contribution function

Assuming a chemical species, X, is emitted from source areas within a possible source region where 3 receptor sites are located. (Number of sites is set to 3 here because only 3 receptor sites were used in our analysis. This method can be applied to other situations where more or less than 3 receptor sites are involved.) After being emitted from a source, a portion of this species, X, undergoes atmospheric reactions and reaches one or more of the three receptor sites in the structural forms of species, $Y_1, \ldots, Y_n$. Another portion of X arrives at one or more of the three receptor sites without undergoing chemical transformation. The possible source region is again subdivided into an i by j gridded array. E_{ij}^{X} is the total emission rate of X from both ground and elevated emission sources, in ton h^{-1}, at each grid cell. The PSCF analysis discussed above offers such results as n_{ij}, which is the number of segment endpoints fall into the ijth cell, and m_{ij}, which is the number of endpoints that correspond to the trajectories that arrived at a receptor site with pollutant concentrations higher than some pre-specified value, and the conditional probability P_{ij} as defined in Eqn. 3 for each species, X and Y_n. The possible maximum amount of species, X, being transported to one of the three receptor sites in the forms of X and Y_n could be expressed as R_{ij}^{x} (ton h^{-1}):

$$R_{ij}^{X} = P_{ij}^{X} E_{ij}^{X} + \sum_{1}^{n} \frac{M_{Y_n}}{M_X} P_{ij}^{Y_n} E_{ij}^{X} \quad (4)$$

where M_X and M_{Y_n} are the molecular weights for X and Y_n, respectively.

The joint PSCF probability for all three receptor sites is needed for the source apportionment of maximum total amount of chemical species being transported to the three receptor sites. Zeng and Hopke [3] used a joint probability, P_{ij}^{T}, for multiple receptor sites calculated by multiplying the conditional probabilities for each single site, i.e., for this particular case,

$$P_{ij}^{T} = P_{ij}^{1} P_{ij}^{2} P_{ij}^{3} \quad (5)$$

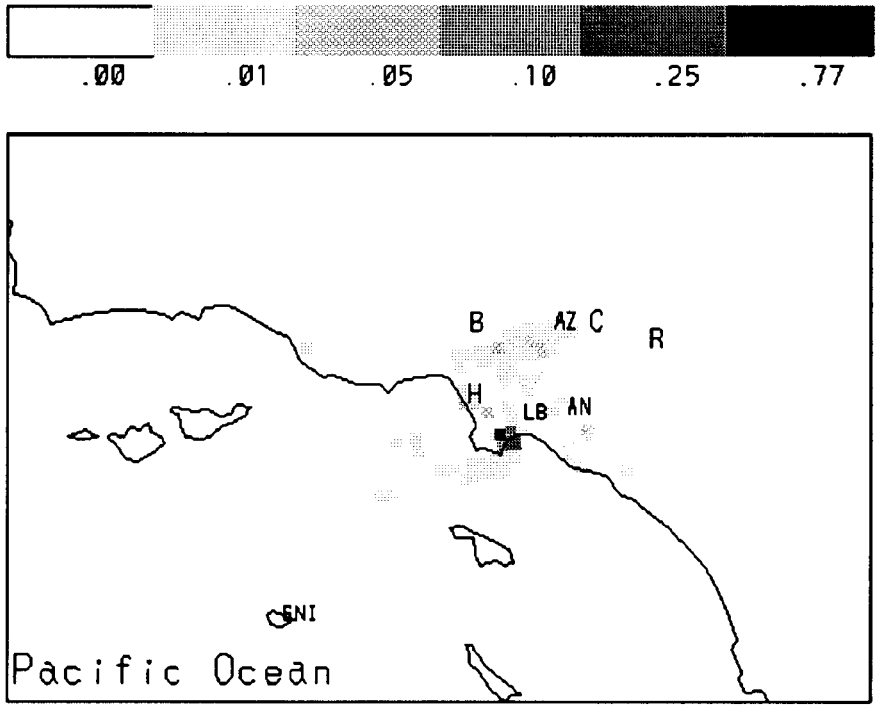


Fig. 12. Source apportionment map of gaseous SO_2 for all three sites, Burbank, Claremont and Rubidoux.

where P_{ij}^1, P_{ij}^2 and P_{ij}^3 are the conditional probabilities for receptor sites 1, 2 and 3, respectively. However, in this particular case involving the Burbank, Claremont, and Rubidoux sites, most of the possible source regions for each of the three sites do not overlap. Thus P_{ij}^T does not exist for most of the grid cells within the studied region. To solve this problem, the calculation for joint probability PSCF for either X or Y_n was modified to yield

$$P_{ij}^T = WT\left[\frac{m_1+m_2+m_3}{n_1+n_2+n_3} + \frac{m_1m_2m_3}{n_1n_2n_3} - \frac{m_1m_2}{n_1n_2} - \frac{m_2m_3}{n_2n_3} - \frac{m_3m_1}{n_3n_1}\right] \quad (6)$$

where m_1, m_2 and m_3 are m_{ij} values for sites 1, 2 and 3, and n_1, n_2 and n_3 are n_{ij} values for sites 1, 2 and 3 from the PSCF analysis, respectively. Given

$$N_{ij} = n_1n_2n_3(m_1+m_2+m_3) \quad (7)$$

a value of 0.5 is given to WT for the case of $N_{ij}=1$, 0.68 for $N_{ij}=2$, 0.85 for $N_{ij}=3$, and 1.00 for $N_{ij}\geqslant 4$. The maximum total amount of species X being transported to all three receptor sites in the forms of X and Y_n is therefore expressed as R_{ij}^T (ton h^{-1}):

$$R_{ij}^T = P_{ij_X}^T E_{ij}^X + \sum_1^n \frac{M_{Y_n}}{M_X} P_{ij_{Y_n}}^T E_{ij}^X \quad (8)$$

The above schemes were applied to the source apportionment of SO_2.

The most important stable product from chemical transformation of SO_2 during atmospheric transport is SO_4^{2-} (SO_4^{2-} PM2.5 was used in the calculation). Therefore, the transport of emitted SO_2 in terms of a single site or all of the three receptor sites are calculated according to:

$$R_{ij}^{SO_2} = P_{ij}^{SO_2}E_{ij}^{SO_2} + \frac{M_{SO_4}}{M_{SO_2}} P_{ij}^{SO_4}E_{ij}^{SO_2} \quad (9)$$

where $P_{ij}^{SO_2}$ is P_{ij} for calculated transport in terms of a single site PSCF, and P_{ij}^T for calculated transport based on the three-site joint PSCF.

The source apportionment results of SO_2 for Rubidoux based on the single site PSCF analysis and the three-site joint PSCF values are shown in Figs. 11 and 12, respectively. The source area near Long Beach contributes SO_2 and SO_4^{2-} at a maximum rate of 0.9 ton h^{-1} to Rubidoux when the meteorology conditions are favorable. It is also seen in Fig. 12 that SO_2 is being brought from off-shore areas to Rubidoux as well as to two other sites. Overall, the maximum contribution of SO_2 emission sources to the three sites is from the Long Beach area at the rate of 0.77 ton h^{-1}, which arrives at those three sites as both SO_2 and SO_4^{2-} (Fig. 12).

Conclusion

Refineries and coastal power plants make identifiable contributions to the SO_2 and nonmarine SO_4^{2-} concentrations observed at the sampling sites. There is also evidence suggesting the contribution from the near-shore regions of the ocean to the concentrations of SO_2 and nonmarine SO_4^{2-}. This may be from ocean shipping. Because of the substantial deposition of sewage sludge into this area, it can also be hypothesized that this organic rich material will cause the sediments to become anoxic and produce reduced gaseous species such as DMS that is then converted into SO_2 and SO_4^{2-}. The combination of emissions data with PSCF results provides quantitative estimates of the area source contributions to the SO_2 and SO_4^{2-} measured concentrations at a given receptor site. However, there is no independent data to permit verification of these results. Further research on the development of this methodology is worthwhile.

This work was supported by the Air Research Board, State of California under a subcontract from the Aero Vironment Inc., California and the National Science Foundation under Grant ATM 8996203. We would like to thank Dr. Chris Pilinis of Aero Vironment who calculated the trajectories and Dr. A.G. Russell of Carnegie-Mellon University for supplying the 1987 emission source inventory data and for his helpful discussions.

REFERENCES

1 G.T. Wolff et al., Atmos. Environ., 25A (1991) 2173–2186.
2 W.C. Malm, C.E. Johnson and J.F. Bresch, in T.G. Pace

(Ed.), Receptor Methods for Source Apportionment, Air Pollution Control Association, Pittsburgh, PA, 1986, pp. 127–148.
3 Y. Zeng and P.K. Hopke, Atmos. Environ., 23 (1989) 1499–1509.
4 A.G. Russell and G.R. Cass, Atmos. Environ., 17 (1983) 949–964.
5 B. Mason, Principles of Geochemistry, New York, 3rd edn., 1966.
6 South Coast Air Quality Management District (SCAQMD), Acid Deposition in the South Coast Air Basin: An Assessment, South Coast Air Quality Management District, El Monte, CA, 1984.
7 W. Bascom, Environ. Sci. Technol., 16 (1982) 226a–237a.

Analytica Chimica Acta, 277 (1993) 381–388
Elsevier Science Publishers B.V., Amsterdam

Environmental applications of combined multidimensional gas chromatography–infrared–mass spectrometry

Kevin A. Krock and Charles L. Wilkins

Department of Chemistry, University of California, Riverside, CA 92521 (USA)

(Received 3rd September 1992)

Abstract

This paper describes the development of a multidimensional gas chromatographic method for the analysis of environmentally interesting samples. In this study, sequential heartcutting across the entire chromatogram of two different paint or lacquer thinner samples was performed. Identification of mixture components was carried out using either computer spectral library search methods or by examination of the infrared and mass spectral data obtained from the analysis. After analyzing the heartcuts, it was determined that the number of identifiable components had increased from 12 to 60 for one thinner sample and from about 60 to 100 peaks for the other sample. These results demonstrate the analytical power of combining multidimensional gas chromatographic separation methods with the qualitative analysis capability of a modern infrared–mass spectral detection system. Some of the difficulties encountered in the present application of multidimensional gas chromatography are discussed.

Keywords: Gas chromatography; Mass spectrometry; Infrared spectrometry; Environmental applications

For a number of years, gas chromatography–mass spectrometry (GC–MS) has been accepted as the method of choice for analysis of volatile environmental samples. More recently, gas chromatography–infrared spectroscopy (GC–IR) and combined GC–IR–MS systems have become accepted as equally useful tools for environmental analysis. The advantages of using a combined GC–IR–MS system over the separated instruments are obvious. First, use of independent spectral detectors provides complementary molecular structure information in addition to retention times, allows low detection limits, and minimizes ambiguities which might arise by performing parallel analysis with different samples using separate analytical systems. However, for maximum utility, it is necessary to deal with a few potential chromatographic problems, inherent in the analysis of complex multi-component mixtures.

One of the main difficulties in gas chromatographic separation of environmental samples is that such samples generally are comprised of extremely complex sample matrices. Use of a single analytical column for analysis of such multi-component samples can compromise chromatographic peak purity. As Davis and Giddings [1] have discussed, for statistical reasons, peak overlap can be a problem. For example, a chromatogram of even an only moderately complex sample separated with a single column is expected to contain no more than about 37% of its potential peaks and only about 18% of the possible pure single-component peaks. Thus, when using a single gas chromatographic separation in a linked GC–IR–MS system, it may be difficult to obtain pure infrared and mass spectra for each component present in the mixture. Obviously, if

Correspondence to: C.L. Wilkins, Department of Chemistry, University of California, Riverside, CA 92521 (USA).

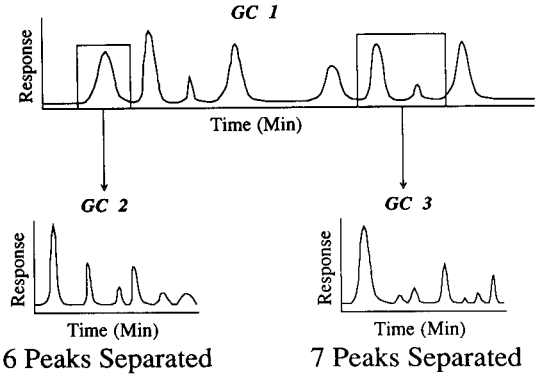


Fig. 1. Theoretical multidimensional gas chromatographic separation. The upper chromatogram is determined from an initial separation on the preliminary column, and the lower chromatograms show the secondary separation on the analytical column.

the resulting spectra are composites arising from one or more components, then use of library search algorithms is compromised and they may not return correct identifications. Thus, for such applications, multidimensional gas chromatographic techniques may be mandated.

The potential advantages of multidimensional gas chromatography (MDGC) have been recognized for many years, since the initial publication by Simmons and Snyder in 1958 [2]. The precise definition of MDGC has been the subject of debate in the literature. However, in the present study, it is defined as the combination of two chromatographic columns of either different or the same selectivity in such a way that eluate fractions from the initial column can be transferred to the second column for subsequent separation. The basic concept is depicted schematically in Fig. 1. In practice, an initial separation is carried out to determine retention time windows for targeted peaks, and in a second (or subsequent) analytical run targeted peaks are cryogenically trapped and then transferred to the second column for separation by a column of the same or different selectivity as the first column. Thus, in the hypothetical example of Fig. 1, the first peak in the GC 1 run is revealed to be comprised of at least six components after the second stage of separation.

Although the majority of gas chromatographic separations are adequately performed using a single column, as Bertsch points out, separations using a single column are dependent on the column efficiency, but in multidimensional systems, the source of the separating power is the difference in selectivity of two or more stationary phases [3]. In addition to the difference in selectivity, another source of separating power for two columns of similar selectivity is the possibility of varying the conditions of the second separation (e.g., oven temperature, column length, stationary phase film thickness, etc.). Another advantage of a multi-column system is the ability to heartcut only a peak (or peaks) of interest from the interfering matrix. While heartcutting techniques are primarily used for targeted compounds, a series of heartcuts can also be taken across an entire chromatogram to enhance the resolution at the expense of analysis time.

Multidimensional gas chromatography has been reviewed extensively [4–7] and many different multidimensional systems discussed [8,9]. The use of MDGC has ranged from hydrocarbon and essential oil analysis to environmental analysis. Examples of recent environmental applications of MDGC include detection and identification of polychlorinated biphenyl congeners by MDGC–MS and MDGC with electron capture detection [10–13] and analysis of glycol ethers in air by MDGC–MS [14]. However, there has only been one reported application of MDGC in an integrated infrared–mass spectrometry system; that application involves identification of components of perfume samples using MDGC–MS–matrix isolation FT-IR [15]. One of the primary advantages of using a combined MDGC–IR–MS system, is its well known potential for differentiating isomeric components [16,17], even if elution order should change between stages of the multidimensional chromatography. Furthermore, for reasons mentioned above, it is precisely in multispectral detector analysis of complex mixtures where MDGC may be most efficacious.

Accordingly, in the present study, analysis of environmentally hazardous components in paint and lacquer thinners by MDGC–IR–MS has been investigated. Such applications are an excellent

test of the analytical concepts involved and provide convincing evidence of the value of employing multidimensional chromatography as part of such analytical protocols.

EXPERIMENTAL

Instrumentation

A commercially available Hewlett-Packard (HP) GC–FT-IR–MS system, modified for MDGC was used to obtain the results reported here. This system includes an HP 5890 Series II gas chromatograph, an HP 5965B Fourier transform infrared detector (IRD), and an HP 5970B mass selective detector (MSD). The IRD and MSD were connected serially using an open split interface. Data was collected with two HP 300 series computers using standard HP IRD and MSD ChemStation software.

Samples

The two samples analyzed in this study were paint or lacquer thinners of unknown origin. The samples were injected neat into the gas chromatograph.

Chromatography

In the MDGC system, both the pre-column and the separating column are placed serially inside the HP 5890 Series II gas chromatograph (see Fig. 2). A J&W Scientific (Folsom, CA) 30 m × 0.32 mm i.d. fused-silica column coated with 5% phenylmethyl polysiloxane (DB-5, 0.25 μm film thickness) served as the pre-column to separate 0.1 μl heated splitless injections of the two thinner samples. The pre-column is connected to a Valco Instruments (Houston, TX) 3-port, 2-way carbon filled polyimide rotor valve which can be switched to route effluent either to the detectors or to the cryogenic liquid nitrogen trap. The cryogenic trap is a 10 cm long, 1/16″ stainless-steel tube with the short end crimped to restrict flow. For injection of trapped heartcuts the tube is warmed by the GC oven temperature to initiate the reinjection onto the analytical column. For the second stage GC analysis, heartcut samples are reinjected onto either a J&W Scientific 60 m × 0.32 mm i.d. DB-5 column with a film thickness of 0.25 μm or a J&W Scientific 60 m × 0.22 mm i.d. DB-Wax (Carbowax 20M) column with a film thickness of 0.15 μm. Heartcuts were taken across the entire chromatograms using both analytical columns. Ultra high purity helium (99.999% He) was used as mobile phase with a head pressure of 70 kPa providing a linear velocity of 33 cm s^{-1} for the initial separation, and after the portion of the chromatogram of interest was heartcut, head pressure was increased to 200 kPa to provide a linear velocity of about 30 cm s^{-1} in the 60 m DB-5 analytical column and about 20 cm s^{-1} in the 60 m DB-Wax column.

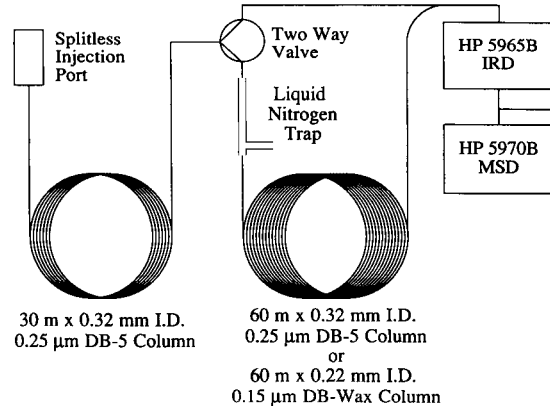


Fig. 2. Schematic diagram of the MDGC–IR–MS equipment used in the present study. Both the preliminary and analytical columns were contained in one oven, along with the liquid nitrogen trap and 2-way valve.

Spectroscopy

Infrared spectra were obtained by using the HP IRD apparatus to collect 10 scans per spectrum with an optical resolution of 16 cm^{-1}. This corresponds to the collection of one spectrum per second. The IRD has a 100-μl (12 cm × 1 mm i.d.) internally gold-coated Pyrex lightpipe that was maintained at 200°C. The detector used in the IRD is a narrow band MCT detector with a spectral cutoff of 750 cm^{-1}. The optics of the IRD are purged with dried nitrogen.

The HP 5970B MSD was run in the full spectrum mode with analysis between 45 and 300 Dalton. Pressures in the MSD throughout all analyses were approximately 3×10^{-5} Torr.

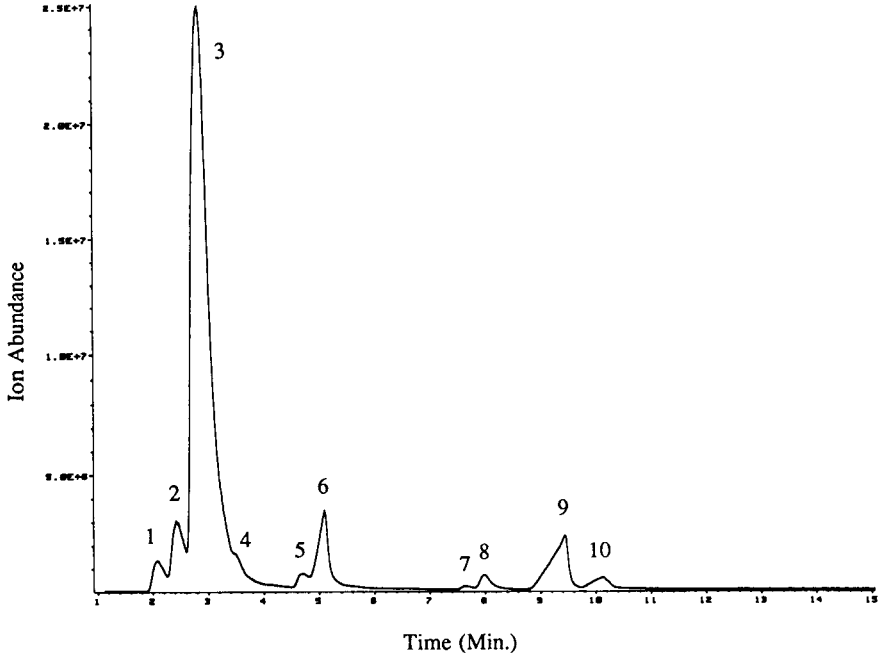


Fig. 3. Total ion chromatogram from the preliminary column for Thinner A. 1 = 2-Propanol; 2 = 2-butanone; 3 = 1,1,1-trichloroethane; 4 = *p*-dioxane; 5 = toluene; 6 = methylpropyl acetate; 7 = ethylbenzene; 8 = *m*-xylene; 9 = 2-heptanone; 10 = 2-butoxyethanol.

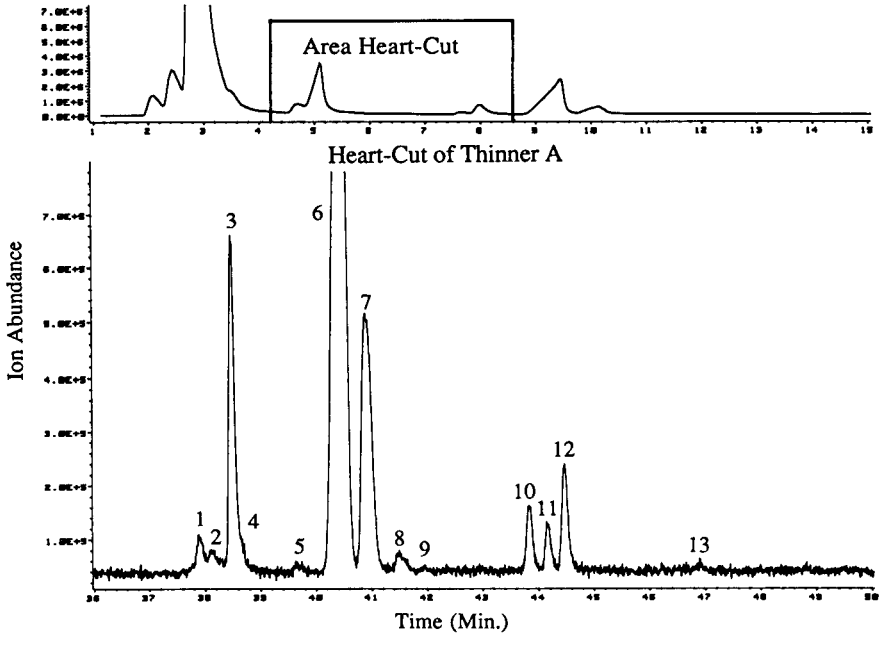


Fig. 4. Heartcut from area shown in top chromatogram for Thinner A. 1 = Heptane; 2 = *cis*-1,3-dimethylcyclohexane; 3 = 1,1,1-trichloroethane; 4 = 2-butanone; 5 = methylpropylketone; 6 = methylpropyl acetate; 7 = toluene; 8 = *p*-dioxane; 9 = butyl acetate; 10 = ethylbenzene; 11 = *p*-xylene; 12 = *m*-xylene; 13 = 2-octanone.

Spectral libraries and searches

The two infrared vapor phase libraries that were utilized in this study were the 3054 spectra HP version of the EPA vapor phase library and the 2004 spectra HP fragrance library. Both libraries were created at 16 cm^{-1} resolution. Spectral library searches were performed using the standard ChemStation search routine with the default search strategy values.

The mass spectral library was the HP version of the NBS library containing approximately 42 000 mass spectra. All searches were performed using the standard probability based matching (PBM) algorithm and standard search strategy values.

When a peak could not be identified by the mass spectral search, but produced library matches with the infrared search, the mass spectrum was used to confirm or reject the infrared hit list members for qualitative identification.

RESULTS AND DISCUSSION

Thinner A and example heartcut

Figure 3 shows the MSD total ion chromatogram from the sample labeled Thinner A. The main component of this is 1,1,1-trichloroethane, which has come under the scrutiny of many environmental chemists as a chemical of particular concern for its deleterious effects upon the atmosphere. The initial one dimensional GC–IR–MS analysis identified nine additional components. However, as Fig. 4 shows, if a heartcut is taken from the area containing the four peaks labeled 5, 6, 7 and 8 in Fig. 3, thirteen components are separated by the 60-m DB-Wax column. Among these thirteen are the four originally identified materials in addition to nine more components resolved from the previously overlapping peaks. 1,1,1-Trichloroethane and *p*-dioxane appear in the heartcut because the cut was made in the tail of the chromatographic peak indicative of overloading by 1,1,1-trichloroethane and including the dioxane shoulder. Also, notice the shifts in the elution order between some of the components. These shifts might not have been detected using a MDGC–MS system, and they would have been even more difficult to detect if a multidimensional GC equipped with thermal conductivity detection, flame ionization detection, electron capture detection, or other non-spectral GC detectors were used.

TABLE 1

Identification of peaks in Thinner A sample [a]

Identification	IR search	MS search	MS confirm
2-Propanol	Y	Y	
2-Butanone	Y	Y	
1,1,1-Trichloroethane	Y	Y	
p-Dioxane	Y		Y
Toluene	Y	Y	
Methylpropyl acetate	Y		Y
Ethylbenzene	Y	Y	
m-Xylene	Y		Y
2-Heptanone	Y		Y
2-Butoxy ethanol	Y	Y	

[a] Y signifies a confirmation by the listed identification technique.

The power of the combined system is clear when examining peaks 10, 11 and 12 in Fig. 4. The complete resolution of the *para* and *meta* isomers of xylene is clearly evident by their infrared spectra, and the presence of ethylbenzene is also obvious from its unique infrared spectrum. All three of these compounds have similar mass spectra, but the difference between ethylbenzene and the xylene isomers is great enough to provide additional support for their identification. Another difficult assignment using mass spectral information alone would be qualitative differentiation between peaks 6 and 9, butyl acetate and methylpropyl acetate, respectively. Here again, the combination of infrared and mass spectra allows positive identification.

Tables 1 and 2 show the component identification for each of the compounds in both the Thinner A sample and the example heartcut. As Table 1 shows, because the PBM search algorithm did not yield positive identification during the first stage of GC analysis, about 40% of the peaks were identified by manual examination of the mass spectrum, even though the peaks appear to be large. This may be caused by the presence of interfering hidden components in the peaks. Sim-

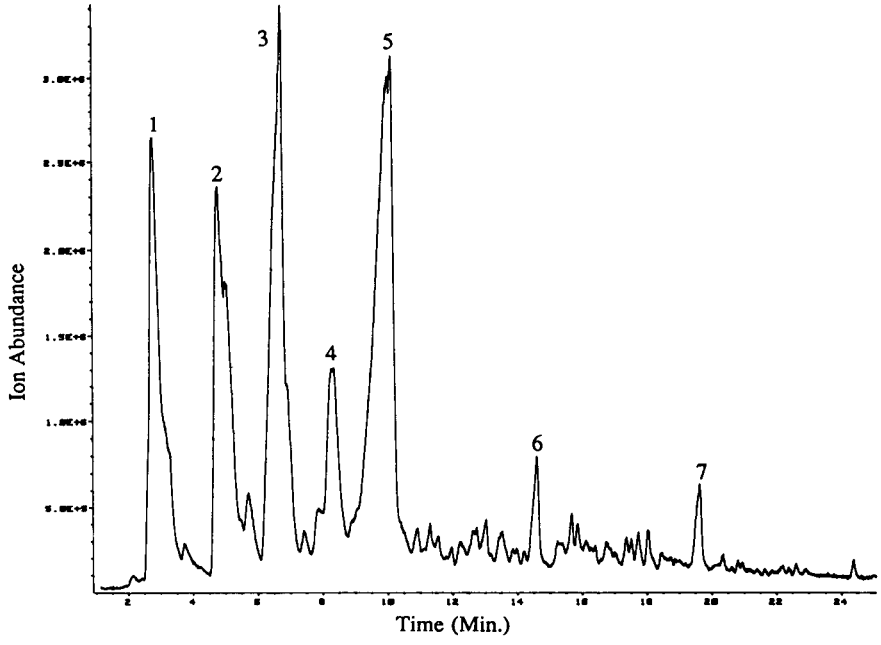


Fig. 5. Total ion chromatogram from the preliminary column for Thinner B. 1 = Isobutanol; 2 = 1-butanol; 3 = methylpropyl acetate; 4 = butyl acetate; 5 = 2-heptanone; 6 = dodecane; 7 = pentadecane.

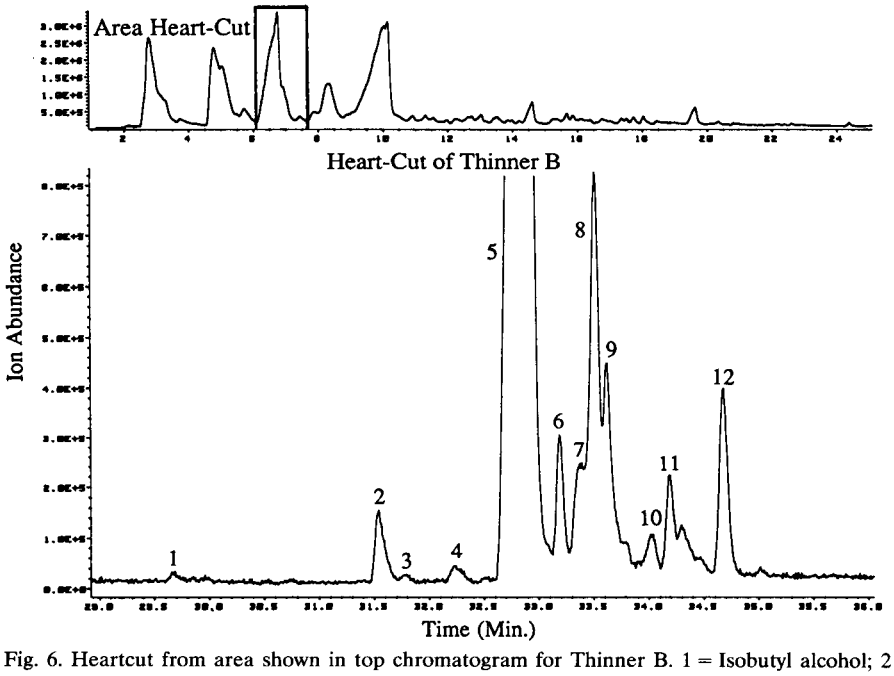


Fig. 6. Heartcut from area shown in top chromatogram for Thinner B. 1 = Isobutyl alcohol; 2 = toluene; 3 = acetate; 4 = octane; 5 = methylpropyl acetate; 6 = 2,6-dimethylheptane; 7 = *trans*-1,2-dimethylcyclohexane; 8 = ethylcyclohexane; 9 = 1,1,3-trimethylcyclohexane; 10 = butylcyclopentane; 11 = 1,3,5-trimethylcyclohexane; 12 = ethylbenzene.

ilarly, Table 2 shows that for heartcut samples about 40% of the peaks were identified in this manner. However, this appears to be due more to increased baseline noise than to the presence of interfering components. Thus, use of the multidimensional separation not only enhances the resolution of the separation, but it also increases the quality of the spectral information obtained and increases reliability of the library search assignments.

Thinner B and example heartcut

In Figs. 5 and 6, the same comparisons are made for Thinner B as for Thinner A, but with second stage GC separation employing a 60-m DB-5 column. In Fig. 5 only seven components are labeled because only the major components were identified to be used as markers for all of the subsequent heartcuts. The heartcut of what appears to be methylpropyl acetate (peak 3), a shoulder, and a small peak in Fig. 5 is revealed by second stage GC analysis to actually be comprised of at least twelve components (Fig. 6). As in the Thinner A example, methylpropyl acetate is present as the main component in the heartcut, and none of the other main components from Fig. 5 appear in Fig. 6, which means that the heartcut was relatively clean and sharp. In Fig. 6, because of incomplete resolution, there are some peaks and shoulders that are not identified. The group of peaks between peaks 11 and 12 include at least 4 different components. It was extremely difficult to obtain clean infrared or mass spectra for these substances. This is an example of the possible need for a third stage of GC analysis, by an additional heartcut of the primary heartcut. Obviously, the number of stages of multidimensional GC required will be a function of the complexity of the sample and the dynamic range required for the analysis.

TABLE 2

Identification of peaks in the Thinner A heartcut example [a]

Identification	IR search	MS search	MS manual confirm
Heptane	Y		Y
cis-1,3-Dimethylcyclohexane	Y		Y
1,1,1-Trichloroethane	Y	Y	
2-Butanone	Y	Y	
Methylpropyl ketone	Y		Y
Methylpropyl acetate	Y	Y	
Toluene	Y	Y	
p-Dioxane	Y	Y	
Butyl acetate	Y		Y
Ethylbenzene	Y	Y	
p-Xylene	Y	Y	
m-Xylene	Y	Y	
2-Octanone	Y		Y

[a] Y signifies confirmation by the listed identification technique.

TABLE 3

Peak number totals for both the single column analysis and the multidimensional analysis for both Thinner A and Thinner B

Sample	Spectroscopic method	Single column peak count	MDGC system peak count
Thinner A	IR	12	61
	MS	12	60
Thinner B	IR	57	87
	MS	60	95

Heartcut summary

Table 3 summarizes the number of components that were separated and identified in the two thinner samples. The numbers represent the number of peaks which could be or have been identified from the single column analysis as well as the total of all of the resulting heartcut peaks. The five-fold increase in the number of identifiable peaks resolved in the Thinner A sample demonstrates the enhanced resolving power of multidimensional gas chromatography. Although the improvement was less dramatic for the Thinner B sample, the number of components identified still increased by almost a factor of two, once again demonstrating the advantage of the MDGC approach.

The selection of the appropriate analytical column polarity is always a difficult task, and for broad heartcuts across an entire chromatogram, there are no obvious guidelines other than to change column polarities depending on the analytes of interest. Late eluting components on a non-polar column tend to be the most non-polar mixture components and there is no way to cleanly

separate non-polar compounds using a polar analytical column. Thus, for components eluting after a certain point in the separation this combination is contra-indicated. However, with a non-polar pre-column and a polar analytical column, up to a point that the system will function with increased resolution, and after that point, the resolution enhancement almost vanishes. With the non-polar/polar system used in this study, retention time was approximately 7 min. All components eluting before 7 min were separated better using the polar column, but all components eluting after 7 min were bunched and impossible to analyze. The only ways around this problem are to use an intermediate polarity analytical column or a column of similar polarity with different properties.

Conclusions

This study has shown that standard one dimensional gas chromatography may not provide the resolving power necessary to fully separate complex samples of environmental interest. It has also shown that use of a multidimensional gas chromatographic approach can be a useful analytical tool for separation and analysis of such samples. Furthermore, the current results demonstrate that application of a simple MDGC approach is compatible with current infrared and mass spectrometric detection systems and can be expected to enhance analytical performance of such instrument combinations.

Support from the National Science Foundation under grant CHE-92-01277 is gratefully acknowledged.

REFERENCES

1 J.M. Davis and J.C. Giddings, Anal. Chem., 55 (1983) 418.
2 M.C. Simmons and L.R. Snyder, Anal. Chem., 30 (1958) 32.
3 W. Bertsch, in H.J. Cortes (Ed.), Multidimensional Chromatography Techniques and Applications (Chromatographic Science Series, Vol. 50), Marcel Dekker, New York and Basel, 1990, p. 74.
4 W. Bertsch, J. High Resolut. Chromatogr. Chromatogr. Commun., 1 (1978) 85.
5 W. Bertsch, J. High Resolut. Chromatogr. Chromatogr. Commun., 1 (1978) 187.
6 W. Bertsch, J. High Resolut. Chromatogr. Chromatogr. Commun., 1 (1978) 289.
7 K. Himberg, E. Sippola and M.L. Riekkola, J. Microcol. Sep., 1 (1989) 271.
8 G.L. Johnson, A. Tipler and D. Crowshaw, J. High Resolut. Chromatogr., 13 (1990) 130.
9 B.M. Gordon, C.E. Rix and M.F. Borgerding, J. Chromatogr. Sci., 23 (1985) 1.
10 D.E. Schulz, G. Petrick and J.C. Duinker, Environ. Sci. Technol., 23 (1989) 852.
11 N. Kannan, G. Petrick, D. Schulz, J.C. Duinker, J. Boon, E. Arnhem and S. Jansen, Chemosphere, 23 (1991) 1055.
12 G. Schomburg, H. Husmann and E. Hübinger, J. High Resolut. Chromatogr. Chromatogr. Commun., 8 (1985) 395.
13 F.R. Guenther, S.N. Chesler and R.E. Rebbert, J. High Resolut. Chromatogr., 12 (1989) 821.
14 E.R. Kennedy, P.F. O'Connor and A.A. Grote, J. Chromatogr., 522 (1990) 303.
15 P.A. Rodriguez, C.L. Eddy, C. Marcott, M.L. Fey and J.M. Anast, J. Microcol. Sep., 3 (1991) 289.
16 C.L. Wilkins, Anal. Chem., 59 (1987) 571A.
17 J.R. Cooper and C.L. Wilkins, Anal. Chem., 61 (1989) 1571.

Analytica Chimica Acta, 277 (1993) 389–404
Elsevier Science Publishers B.V., Amsterdam

Atmospheric chemometrics for identification of trace element sources in precipitation

Richard J. Vong

College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331–2209 (USA)

(Received 3rd September 1992)

Abstract

This paper demonstrates solutions to some of the problems normally encountered in the analysis of atmospheric chemical data sets. Multivariate data analysis techniques were applied to a unique chemical data set consisting of rainwater trace element concentrations determined via instrumental neutron activation analysis (INAA). Experimental uncertainties, principal component analysis (PCA) and principal component classification were used to categorize sources of variability in the data while partial-least-squares regression (PLS) was used to compare the rainwater data to crustal, sea salt, and smelter "signatures" of the same trace elements. These three source contributions to the rainwater trace element data were calculated and four other contributions from derived sources were interpreted based upon their chemical composition, geographical distribution, and temporal variation.

Keywords: Neutron activation methods; Principal components analysis; Chemometrics; Precipitation; Rainwater

Over the last two decades atmospheric chemists have presented analyses of aerosol, gas, precipitation, and cloudwater data that include univariate and multivariate treatments intended to address problems such as identification of emission source [1,2] and meteorological influences [3], blank corrections [4], and the effect of experimental uncertainty [5]. However, for many investigators in the field of atmospheric chemistry there remains a certain reluctance to employ statistical techniques such as principal component analysis (PCA) or multivariate regression because these techniques sometimes are regarded as "black boxes" which cannot describe a real world physical/chemical system such as the atmosphere. At the same time, a wealth of data on the chemical state of the atmosphere has become increasingly available as a result of the application of multielemental techniques such as instrumental neutron activation analysis (INAA), proton induced x-ray emission spectroscopy (PIXE), inductively coupled plasma emission spectroscopy (ICP), ion chromatography (IC), and x-ray fluorescence (XRF). This paper is intended to illustrate a number of approaches to such a multivariate chemical data set, and, thus, help familiarize atmospheric chemists with the field now known as "chemometrics".

Chemometric techniques have been applied to a unique data set consisting of the chemical composition of rainwater collected by a 38 site network in the Puget Sound area of Washington State, USA. Rainwater samples were collected both before and after the permanent closure of a large copper smelter. The experimental design produced both spatial and temporal information as well as the measured experimental uncertainty. Temporal information was obtained by analyzing

Correspondence to: R.J. Vong, College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331–2209 (USA).

six precipitation events, especially by contrasting three events sampled before the closure of a major emission source (a copper smelter) to three events sampled after that source closed. Spatial information was used to interpret whether an influence was due to a particular industrial source (coherent spatial pattern) or rather to a more random process. Experimental uncertainty was used to select the variables for the data analysis and to aid in the data interpretation.

In previously reported results from this field experiment, the smelter's contribution to precipitation sulfate, nitrate and hydrogen ion was estimated by comparing upwind and downwind ion concentrations using an analysis of variance (ANOVA) model [6]. For the present analysis, the primary chemical data consist of both soluble and insoluble trace element concentrations in rainwater as determined via instrumental neutron activation analysis (INAA). Maps for a few trace elements (As, Al) have been presented for two individual storms [7,8].

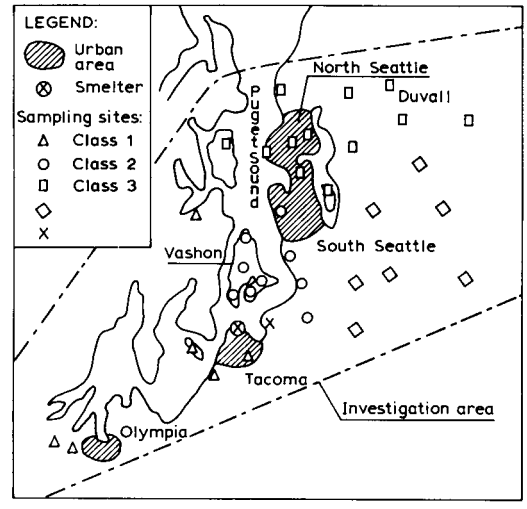


Fig. 1. Map of western Washington State, USA, with the smelter, rainwater sampling sites, as well as urban and geographical areas indicated.

EXPERIMENTAL

Each sampled storm was preselected according to a criteria that specified winter rains from cyclonic frontal systems with south–southwesterly winds and relatively uniform precipitation over the network. Sites were located in urban and rural areas both upwind and downwind of the major air pollution source in the Northwestern USA, a copper smelter. Figure 1 displays a map of the Puget Sound area indicating the location of the sampling sites, the copper smelter, and urban areas of western Washington State (Seattle, Tacoma, and Olympia).

The sampler consisted of a polyethylene funnel and bottle. The samplers were prewashed in HNO_3 and repeatedly rinsed in deionized water until the final rinse conductivity was < 1 μS cm^{-1}. Random chemical analysis of the rinses indicated no detectible contamination. For two storms, four to ten collectors were co-located at a representative site and the collected precipitation was chemically characterized in order to generate information about overall (sampling and chemical analysis) experimental precision.

The rainwater samples were returned to the laboratory for pH measurement and filtration (0.2 μm pore size) prior to frozen storage. For about 150 samples, the filtrate and the filters were analyzed separately by instrumental neutron activation analysis (INAA). 50 ml of the filtrate was acidified and freeze-dried to residue prior to INAA; soluble species are defined from these analyses. The insoluble fraction is defined as the material that was retained on the filters [9]. Fourteen storms were sampled by the network but only six were characterized by full INAA analyses.

INAA (for Al, Ca, Ti, V, Cu, Mn, Dy, Na, K, As, Br, Sb, La, Sm, Mg, Fe, and Se) was performed at Los Alamos National Laboratories Omega West reactor using thermal neutrons [10]. Samples were irradiated twice and counted three times after appropriate decay periods. The first irradiation ("shorts") was for 5 min with a flux of 6×10^{12} $n/cm^2/s$ followed later by a second irradiation ("longs") for 4 h with a flux of 9.7×10^{12} $n/cm^2/s$. "Shorts" were counted for 5 and 19 min (after 5 or 20 min decays, respectively) on Ge(Li) detectors. "Longs" were counted on Ge detectors for 4 h (after a 4 day decay). A second

TABLE 1

Gamma ray lines used for INAA (in keV)

Short irradiation species			
Al	1779.0		
Ca	3084.15	4071.0	
Mg	1014.24		
Ti	320.08		
V	1434.05		
Cu	1039.0		
Na	1368.6	2753.85	
Mn	1810.66	2113.0	
Dy	94.68	279.74	361.66
Long irradiation species			
Na	1368.5	2753.9	
K	1524.7		
As	559.5	657.2	1216.3
Br	776.5	554.3	619.1
Sb	564.1	692.8	
La	487.1	815.7	1596.4
Sm	103.2	69.66	
Se	264.7	279.4	
Fe	1099.2	1291.6	

count of the "longs" was sometimes performed. The detectors were calibrated using NIST (NBS) reference materials for coal, coal fly ash, urban particulate, and orchard leaves; a multielemental analysis program ("Raygun") was used for identification and quantification of peaks [11]. Table 1 describes the gamma ray lines used to identify and quantify the INAA trace elements. Cobalt flux detectors were used to correct the neutron fluxes for sample geometry. Further INAA analytical details are available elsewhere [9].

DATA ANALYSIS METHODOLOGY

Blanks and experimental errors

The mean of a number (N) of blank analyses was compiled for each analyte and sample type. A limit of detection (LD) was defined [11] from the standard deviation of the blank and N. Elements with blank corrected masses between zero and LD were considered to be undetected, retained in the data base, and filled by a random number between zero and LD to preserve information that the mass was low. The insoluble and soluble concentrations from INAA were determined from the blank corrected masses and either the filtered or freeze-dried rain volume. The precision of the various INAA species was between 5 and 30% of the measured concentrations, depending on species and absolute concentration. To assure the intercomparability of the chemical data obtained via different techniques, we compared results for Na in the soluble phase because it could be precisely determined from three (ICP, IC, and INAA) techniques. In addition we compared the INAA "shorts" and "longs" results for Na. These comparisons indicate very good agreement ($0.93 \leq r^2 \leq 0.97$). Table 2 summarizes the experimental uncertainties for the trace elements in rainwater determined via INAA. Besides the Al, Ca, Ti, V, Cu, Mn, Dy, Na, K, As, Br, Sb, La, Sm, Mg, Fe, and Se data presented here, several other species were occasionally detected by INAA but are not included here either because of large experimental uncertainties (I; Cl; high blank values in combination with expected volatilization) or a low frequency of detection (In, Zn, Cd, Eu, Au, Ag, Cr, W, Ga).

Data transformation and normalization

It is customary to mean center and variance scale data (thus producing z-scores) prior to PCA. The variance scaling allows each variable (column) the same opportunity to participate in the principal components since each will have a variance of 1.0 after scaling. An additional scaling termed "row normalization" sometimes is applied prior to the above steps. This approach divides every sample concentration by a term that can be thought of as the same sample's "strength" in order to reduce covariance in the data that is unrelated to source contributions. Typically an atmospheric chemical data set that has not been "row normalized" will display covariance between many or all chemical variables regardless of independent emission source influences [12] due to meteorological variation (e.g., dispersion). After normalization the data describe whether the elemental concentrations are a larger or smaller fraction of the total concentration (sample strength) than average. Normalization of these data eliminated any need for additonal transformation (e.g., log or square root) because the

TABLE 2

Blanks values, detection limits and measured experimental precision for instrumental neutron activation analysis of rainwater samples

Species	Blank value (ng sample^{-1}) [a]	N [b]	Detection limit [c] (pg l^{-1})	Precision [d] (% conc.) [e]
Soluble fraction				
Short irradiation species				
Al	1800	34	1000	3– 6
Ca	nd [f]	0	–	8
Mg	nd	0	–	–
Ti	nd	0	–	–
V	16	34	12	4– 7
Cu	nd	0	–	21
Na	2700	30	5000	4– 6
Mn	9	10	35	9–10
Dy	nd	0	–	–
Long irradiation species				
Na	2900	14	11 000	4– 5
K	nd	0	–	–
As	< 1	1	< 30	7–13
Br	11	10	70	9
Sb	0.7	14	4	9–11
La	1.5	9	7	–
Sm	0.2	10	1	33
Se, Fe	nd	0	–	–
Insoluble fraction				
Short irradiation species				
Al	500	20	210	14–31
Ca	nd	0	–	22
Mg	nd	0	–	–
Ti	nd	0	–	16–36
V	5	22	3	20–28
Cu	nd	0	–	28–29
Na	2100	21	1600	12–24
Mn	< 7	0	< 100	22–31
Dy	nd	0	–	33
Long irradiation species				
Na	2100	14	1800	14–24
K	700	9	400	4
As	nd	0	< 10	28–34
Br	25	11	20	9–10
Sb	0.5	11	1	23–24
La	nd	0	–	12–20
Sm	0.1	13	< 1	8–21
Se, Fe	nd	0	–	–

[a] Mass detected for one blank filter plus the sample bag. [b] N = Number of blanks where a given species was detected. [c] Aqueous detection limit (DL) after blank subtraction [11]: $DL \equiv [ts/N^{0.5}]/v$. where: v = sample volume (soluble = 50 ml, insoluble = 200 ml), s = blank standard deviation for N replicates, t = student's t-statistic for $(N-1)$ degrees freedom at 99% CI. [d] Experimental precision includes sampling and chemical analysis errors and is based on colocated sample collection at one site for two storms. N samples = 9 per storm for "short" irradiation. N samples = 4 or 6 per storm for "long" irradiation. [e] Coefficient of variation (S.D./mean) expressed as percent; range of values for two storms; one additional storm at a different site also was used for determining the insoluble fraction uncertainties for "short" irradiation species. [f] nd = Not detected.

normalized data were approximately normally distributed.

Such row normalization procedures can produce spurious correlations in the data when one or a few elements dominate the normalization factor (sample strength); this problem is termed "closure" [13]. To avoid this closure problem one can adopt the approach taken by the Lund (Sweden) group [14,15]. For the present application, the sample strength (normalization factor) is calculated from the concentration of five elements that are well determined experimentally and of interest environmentally (Al, V, Mn, As, Na). The elements are each expressed as the ratio of the sample value to the average for all samples before they are added to form the normalization factor. In this manner the sample strength is derived, on the average, from five equally weighted elements in accordance with the recommendation of Johansson et al. [13]. A recent study of atmospheric aerosol obtained comparable PCA results on row normalized data based alternatively on: (a) an independently determined gravimetric mass as sample strength and (b) derived 'sample strength' as above [16]. It is straightforward to back-transform after such a normalization by using each sample's "strength" to convert data and the PLS predictions to original concentration units.

Effects of solubility

In a separate analysis to be presented elsewhere, the soluble/insoluble partitioning was investigated based on the separate INAA analyses of the solute and filtrate. Those results address the question of how trace element solubility affects the correlations among species. Insoluble Al comprised about 90% of total Al. For As, Sb, and Cu the two phases were correlated with about two-thirds in the soluble phase. Na and Br were mostly soluble. These species, by virtue of most of the mass being present in one phase or due to the correlation between the insoluble and soluble concentrations, produce similar conclusions as to the type of source contributions regardless of whether insoluble and soluble concentrations are used as separate variables or are added together to form "total concentrations". Non-linear partitioning between the phases as well as uncertainty as to the solubility of most anthropogenic emissions (probably a function of the air's history) suggests that the conservative "total concentration" is the appropriate metric for the evaluation of source mass contributions.

Geographic classes

I was able to identify specific regions within the sampling network that had relatively homogeneous emission source influences because of: (a) knowledge of the locations of the largest air polution sources in Washington State, (b) experimental simplifications associated with meteorology (wind direction which was known and consistent; "clean" background air exists to the west of the network [5]), and (c) a previous analysis (using ANOVA) of SO_4^{2-} and H^+ data from the same experiment [6]. The regions are identified as classes 1, 2, and 3 in the principal component classification for upwind, "smelter-influenced", and urban/rural areas, respectively.

Multivariate analysis techniques

Typically, multivariate data analysis techniques are useful for data sets consisting of many measureable characteristics in order to reduce the dimensionality (number of variables that must be interpreted) to a more manageable set of new variables ("components"). In many cases these components can be analyzed (post-processed) using familiar geophysical and/or graphical approaches, thus putting physical scientists into their "comfort zone" while at the same time retaining much of the information inherent in multielemental chemical analyses. The present analysis relies heavily on features of the experimental design to facilitate this post-processing of the predictions.

PCA [1,17,18], principal component classification–"SIMCA" [19], and PLS regression [20] techniques that were applied to the present data set have been fully described elsewhere. Briefly, both PCA and PLS extract factors which describe the variance in a data matrix (two matrices for PLS) which typically consists of rows that represent samples (time or space variation) and columns that represent chemical measurements

performed on those samples. Orthogonal PCA or PLS components are successively extracted from the data in order to reproduce the systematic variation in the original data until appropriate stopping rules [21] indicate that only random variation remains. "Loadings" characterize the correlation between the original chemical concentrations and the components while "scores" describe the sample-to-sample variation in the components (thus characterizing the strength of the influence whose chemical makeup is described by the loadings). In the present work the PCA and PLS components were obtained using the NIPALS algorithm developed by Wold [22].

In PLS components are simultaneously calculated for two matrices (**X** and **Y**) with the scores in **X** related to the scores in **Y** through a regression coefficient. The PLS **X** components are similar to those that PCA would extract from the same matrix except that in PLS the **X** components are, in effect, rotated to maximize their ability to predict **Y**. The PLS components are orthogonal; the term "rotation" refers to the exchange of information (scores) between **X** and **Y** during the original calculation of the PLS components [20].

A special form of partial least squares regression termed "discriminant PLS" was used in the present study; here **Y** is a "design" or classification matrix rather than additional chemical or physical information. In the present study a model "training" data set consisting of source chemical profiles was used to predict one **Y** variable for each independent source influence [given N "pure" sources, source a has a value for $y(a, a)$ of one in the PLS training model while all other $y(a, i)$ are set to zero]; if available, multiple profiles can be used to estimate each source influence and view the associated uncertainty. Colinearity is not a problem with PLS modeling because it extracts components from the original data but contributions from colinear sources cannot be independently evaluated. However, the training set must contain profiles for all the possible source influences or its application to data consisting of mixtures will produce spurious results (e.g., mixing fractions $\gg 1$ or $\ll 0$). The application of discriminant PLS to a synthetic data set where the source contributions are known has been previously reported [23].

When the PLS model is used with a data set consisting of mixtures of the same sources (e.g., the rainwater samples), each **Y** prediction represents the mixing fractions of the "pure" sources (training set). The sum of all mixing fractions is forced to 1.0 by the appropriate selection of the "design" **Y** matrix. Mass concentrations are obtained by multiplying these mixing fractions by the normalized (discussed above) source profiles and the normalization factor for each rainwater sample.

RESULTS

A series of multivariate analyses, each with a particular set of objectives, was performed on the INAA data using PCA and PLS. To a certain extent one refines the original objectives of successive PCA and PLS analyses in order to maximize the usefulness of the information obtained. Inherent in multiple analyses is the possibility that one uses up unaccounted for "degrees of freedom" by tailoring the later analyses to insights gained from earlier steps. Such 'data snooping' is most likely to represent a problem when strict significance levels are to attributed to various results of the analyses. In all cases the subsequent analyses should be based on objective criteria that are not derived from earlier, and related, data analyses (e.g., in PCA one should not stratify data into classes based a criteria derived from a previous PCA).

Mean composition and stoichiometry

The mean trace element composition of 125 rainwater samples is presented in Table 3 along with seven chemical influences on these samples (discussed below). Consideration of sample stoichiometry, air pollution emission inventories, and the geographical distribution of individual trace element concentrations suggests at least three source contributions to precipitation composition in the sampling region. The geographical distribution of As concentration for two precipitation events demonstrated clear maxima immediately

downwind of the copper smelter (the largest source of airborne As in the USA prior to closure), as expected [7,8]. Three species (Na, Br, and Mg; also Cl from IC analyses reported in Ref. 6) are present in ionic ratios comparable to sea salt [24]. A number of elements (Al, Fe, Mn, Sm, La, Ti, and Dy) were present, on the average, in mass ratios fairly typical of the earth's crust [9,25].

Thus, the multivariate analyses were, in part, designed to: (a) resolve the contributions of the copper smelter, sea salt, and crustal material to rainwater composition and (b) identify other (less obvious) contributions. It is likely that other emission sources also influenced the rainwater chemical composition. A previous analysis suggested that part of the sampling network received substantial sulfur contributions from sources other than the copper smelter [6].

The trace element composition of the smelter emissions is presented in Table 3; the source profiles for the smelter were constructed from aircraft sampling of the plume during one of our rain events [26] and by stack gas sampling [27] that was performed several years prior to our experiment. Three profiles were constructed to reflect uncertainty in the As/Sb/Fe ratios in the smelter emissions. Also presented in Table 3 are the composition of sea salt and average crustal material as well as four derived "other influences" that will be discussed below.

PCA modeling

A number of multivariate data analyses were performed using several different (complementary) approaches. The first approach utilized separate soluble and insoluble concentrations (results for data without row normalization are discussed here) for eighteen trace elements from INAA analyses (Al, Ti, Ca, V, Mn, As, Br, Sb, Na-'shorts', Na-'longs', Mg, K, Fe, Se, Cu, La, Sm, Dy) of 131 rainwater samples. This PCA

TABLE 3

Chemical composition of crustal (1), sea salt (2) and smelter (3) source profiles, four groups of samples used to describe "other influences" and rainwater (125 sample mean)

ID No.	Chemical composition																
	Al	Ti	Ca	V	Mn	As	Br	Sb	Na	Mg	K	Fe	Se	Cu	La	Sm	Dy
Crust																	
1	1	0.07	0.50	0.002	0.012	0	0	0	0.29	0.28	0.25	0.68	0	0.001	0.0004	0.0001	0.00004
Sea salt																	
2	0	0	0.04	0	0	0	0.006	0	1	0.12	0.04	0	0	0	0	0	0
Smelter																	
3	0	0	0	0	0	1	0	0.16	0	0	0	0.15	0.5	0.4	0	0	0
	0	0	0	0	0	1	0	0.12	0	0	0	0.09	0.5	0.17	0	0	0
	0	0	0	0	0	1	0	0.12	0	0	0	0	0.5	0.17	0	0	0
	Rainwater (ng l^{-1})																
	Al	Ti	Ca	V	Mn	As	Br	Sb	Na	Mg	K	Fe	Se	Cu	La	Sm	Dy
Mean																	
	127	14	173	0.44	3.8	2.4	2.0	0.27	1160	280	196	99	1.7	6.7	0.054	0.009	0.15
Other influences (see text)																	
	Al	Ti	Ca	V	Mn	As	Br	Sb	Na	Mg	K	Fe	Se	Cu	La	Sm	Dy
4	37	8	42	0.23	0.45	0.45	0.8	0.05	400	–	32	22	–	–	0.015	0.001	0.039
5	86	8	50	0.33	1.6	0.13	0.6	0.03	300	38	50	70	–	–	–	0.002	0.014
6	20	9	228	0.12	2.5	0.06	3.8	0.02	1700	160	470	48	–	–	0.013	0.003	0.014
7	48	10	40	0.12	1.2	0.4	0.37	0.015	50	–	48	45	–	–	0.05	0.200	0.006

extracted three significant components (based on cross-validation, Ref. 21) that characterized insoluble "crustal elements", soluble "sea salt" elements, and a factor related to As, Sb, Cu and Se concentrations in both soluble and insoluble phases, respectively. These three dominant sources of variability (principal components 1–3) represented 63% of the original data variance.

2-D projections of PCA "scores" represent a very powerful and a quick means of identifying unusual samples. Such samples might represent errors but in certain cases these "outliers" might be particularly valuable samples (prime candidates for detailed case studies). This 'first' PCA identified three samples as clear "outliers" (score plots revealed these samples to be far from the bulk of the data; further inspection revealed unusual soluble K or insoluble Br concentrations). These "outliers" were discarded so that subsequent analyses would reflect the variation in the bulk of the samples (128 samples remaining).

Classification

For the 'second' PCA in this series of analyses, the trace element data were stratified a priori into three regions ("classes") that should have received background air, "smelter-influenced" air, and urban/rural air (that is characteristic of a 40 km by 40 km area in and near Seattle and Tacoma, but not in an area that was demonstrated to be "smelter-influenced"). Principal component models were constructed for each class on variance-scaled (separate soluble and insoluble variables) rainwater chemical data that had not been row normalized. These class models were compared by (a) determining which chemical species distinguish the three geographic areas based on their modeling power [19,28,29] (b) calculating confidence intervals around the scores for each class in order to determine whether the classes were distinct or similar in terms of trace element composition.

Each sample was fit to each class model and

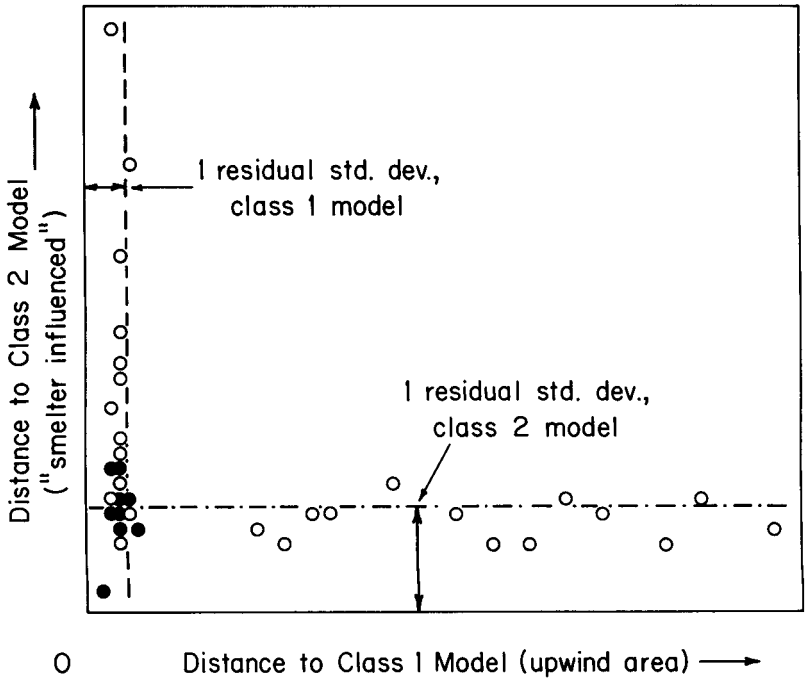


Fig. 2. Residual standard deviations for upwind (class 1) and Vashon (class 2) samples fitted to the class 1 and 2 models. The lines indicate one residual standard deviation for samples fitted to their "own" class model; filled circles represent multiple samples.

the corresponding "distance" between the sample's location in N-dimensional space (where N is the number of chemical species) and the principal component class models (a one dimensional principal component model is a line; a two-dimensional (component) model is a plane, etc.) was determined. A sample was implied to be a member of a class or not ("similar" to the class in that it is not significantly different from that class model) by comparing the residual obtained ("distance") when that sample was fit to the class model to the within-class residual variance. Formula for these calculations are available elsewhere [19,28,29].

Figure 2 presents a plot of these "distances" for the samples collected in the upwind area (class 1) of the network versus those collected immediately downwind of the smelter (class 2; previously identified as "smelter influenced" in Ref. 6). The class 2 samples are clearly different than the class 1 samples as their "distances" to the class 1 model all exceed twice the residual standard deviation of the class one model. Most, but not all of the class 1 samples are different from the class 2 model (a few, including overlapping samples) lie within one residual standard deviation (horizontal line) of the class 2 model. These samples have a chemical composition that falls in a region where the two class models overlap (bottom left corner of Fig. 2). Similar results were obtained when class 2 was compared to samples collected out of the region of obvious smelter influence (class 3; that is the Seattle urban area, the Duvall rural area, and the Vashon area for the year when the smelter was shut down). A similar comparison of the upwind (class 1) samples with the class 3 (Seattle-Duvall) samples demonstrated a large degree of overlap; many upwind rainwater samples are not significantly different than some of the samples collected in the Seattle-Duvall area (especially those collected after the smelter closed).

To assess the overall degree that the classes were distinct, we calculated the class separations [28,29] from the ratio obtained from the variance of the out-of-class samples (sum of squared dis-

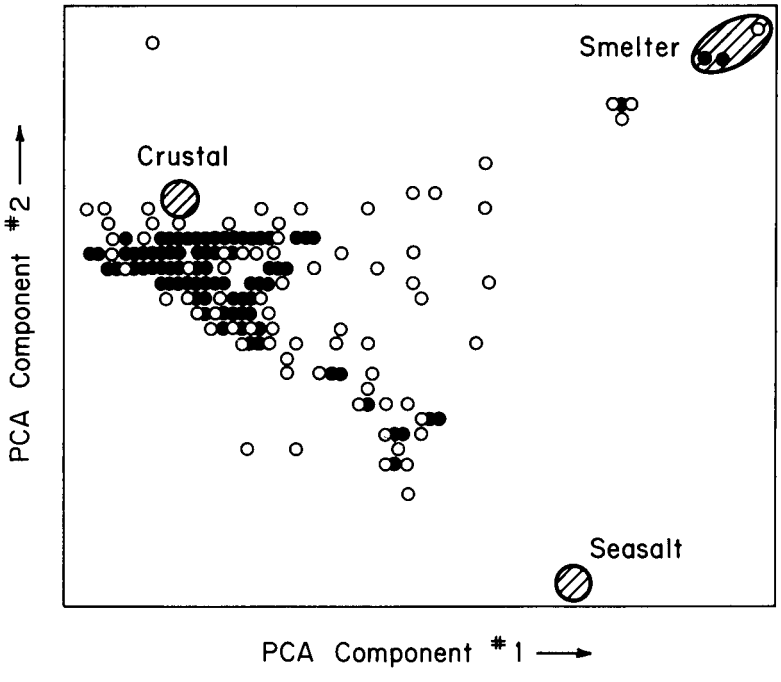


Fig. 3. PCA components 1 and 2 scores for rainwater samples and three types of source profiles (crust, sea salt and three smelter signatures).

tances) fitted to the class model divided by the variance from the within-class samples fitted to their own class model (class residual standard deviation) for the three classes. These differences were largest when class 2 was compared to the others. The only differences that were not clearly significant ($p < 0.01$) were when class 1 samples were fit to the class 3 model; the residual standard deviation of the class 3 model was somewhat larger than for classes 1 and 2. The variables which had the greatest modeling power were soluble Na for class 1, soluble and insoluble As and Sb for class 2, and insoluble Al for class 3.

Source contributions

A discriminant PLS model was used to calculate the contributions of sea salt, crustal material, and the copper smelter to individual rainwater samples. Twelve PLS components were extracted from the five profiles (three similar ones for the smelter) and ten rainwater samples (four types of "other influences", see below) thus decribing the maximum dimensionality for a "closed" data set with 13 independent influences. Each component was significant according to a cross-validation criteria [21]. Figure 3 displays the three source types along with the rainwater samples as a score plot from the previous PCA. Since all mixtures consisting only of the crustal, sea salt, and smelter emissions must fall interior to the triangle that these three sources form [23], it is evident that additional source influences exist. While it is possible to postulate specific sources (e.g., wood smoke, coal ash, oil combustion, automobile exhaust, road dust) and use available "library" emission source profiles [2,18], in the present study I have chosen to derive information regarding these "other influences" (i.e., other than sea salt, crust, and smelter).

The rationale for the decision not to use library source "fingerprints" is that: (a) the library profiles may not describe Washington State coal, oil or wood emissions, (b) unknown sources may still be present once profiles are selected, (c) some chemical species (noteably Pb, C, and Ni) that distinguish these source types were not avail-

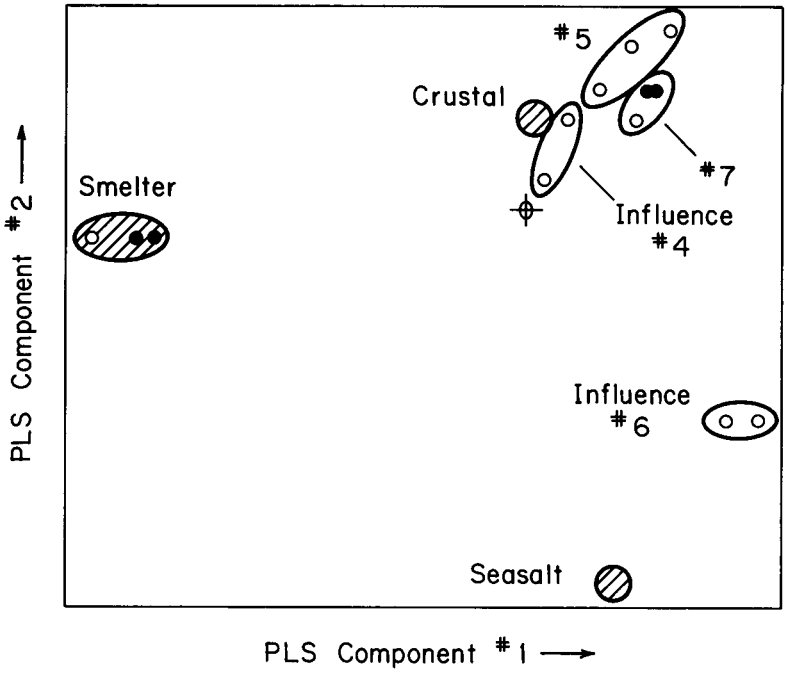


Fig. 4. PLS training model scores for components 1 and 2 for the source profiles and "other influences" (identified as: crust, sea salt and three smelter signatures; groups labeled 4, 5, 6 and 7 are the "other influences" described further in Table 3).

able from INAA, and (d) some of these profiles are quite similar to crustal material (e.g., coal) making it difficult to calculate the associated source contributions independently [18,23].

Four "other influences" were identified by inspection of "scores" plots similar to Fig. 3 (for combinations of components 1–8) and selecting the samples that would define a new "volume" or "mixing space" that, along with the three source profiles, encloses the remaining samples. Each of the four new vertices was characterized by two or three similar rainwater samples. This new seven sided hypercube (three sources and four samples) replaced the triangle of possible mixtures formed from the crust, sea salt, and smelter profiles. Figures 4 and 5 present PLS **X** scores; these plots illustrate the revised mixing space. Note that the differences among the three smelter profiles has little effect on the mixing space while the four "other influences" can be quite important to defining the mixtures that the discriminant PLS model can separate (depending on which lower dimensional 'scores' projection is viewed, of course).

Figures 4 and 5 (and other projections) indicate that the four "other influences" are similar to crustal material (and in the case of 6, sea salt as well). While loadings plots (not shown) provide information as to the differences between the average crustal material (1) and influences 4, 5, 6, and 7, I have chosen to evaluate these "sources" using the enrichment ratio technique that is familiar to atmospheric chemists [5,9]. The ratio of an element to Al in the samples is compared to that same ratio for crust to determine which trace elements are "enriched" relative to the ratios in crustal material. Table 4 presents the results of this approach for each of the "other influences".

It is evident from the PLS projections (Figs. 4 and 5) and the enrichment factors (Table 4) that influences 4–7 are crustal-like material with higher ratios of Ti, Ca, Dy, Sm, La, K, Mg, Na to Al than found in the average crust. In addition As and Sb are highly enriched in three of these

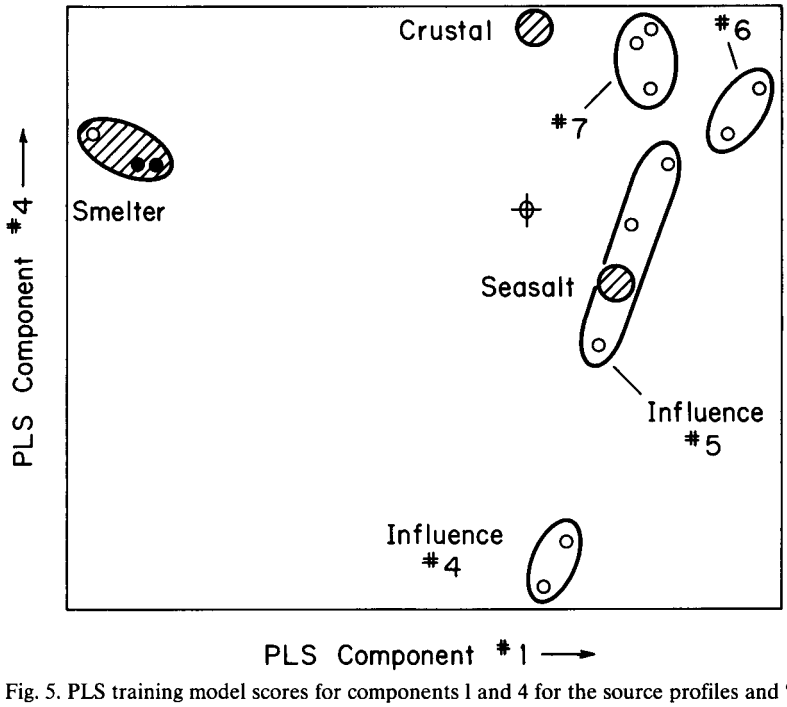


Fig. 5. PLS training model scores for components 1 and 4 for the source profiles and "other influences" (identified as: crust, sea salt and three smelter signatures; groups labeled 4, 5, 6 and 7 are the "other influences" described further in Table 3).

TABLE 4

Enrichment factors for comparison of the chemical composition of four "other influences" (4–7) on rainwater composition (those derived from ten unusual samples the data set) with crustal and sea salt elemental ratios

[The elemental ratios in sample chemical composition for each of these "other influences" (each ID No.) were compared to the same elemental ratios for crustal (ratio to Al) and sea salt (ratio to Na) material. The elements listed have ratios to Al in the selected rainwater samples at least three times as high as could be explained by "crustal material" (enrichment factor = 3)]

ID No.	Species	Crustal enrichment ratio (relative to Al)	Sea salt enrichment ratio (relative to Na)
4	As, Sb	300	[a]
	Ti, Ca, V, K	4	[a]
5	As, Sb	30	[a]
	Na, Br	12	[a]
	Dy	5	[a]
6	Na	250	
	Br	> 250	0.4
	Mg	25	1.0
	Ca	22	4
	Dy	15	[a]
	Mn, Ti	10	[a]
	K	10	8
	Fe	4	[a]
7	As	300	[a]
	Sm, La	50	[a]
	Sb	10	[a]
	Na, Br, K	4	[a]
	Ti, Mn, Dy	3	[a]

[a] These sea salt enrichment ratios were not determined. Most of these elements have very low concentrations in sea water. Br sea salt enrichment values of less than 1.0 could result from Br and Cl volatilization on the surface of acidic aerosol.

TABLE 5

Rainwater mixing fractions for crustal (1), sea salt (2), smelter (3), and four "other (derived) contributions" for different geographical regions [a,b] and years

[The grand average sum of the mixing fractions for any data sub-set (the eight region year classes) would be 5 (the five normalizing elements for sample strength, Al, V, Mn, As and Na, would each contribute a value of 1.0 when its concentration was at the mean); classes that sum to > 5 have rainwater Al, V, Mn, As, and Na concentrations (their sum) greater than average]

Regions [a,b]	Year	Relative mass source contributions							*N*	Sum of all (% of mean)
		1	2	3	4	5	6	7		
Upwind	1985 [c]	0.16	0.96	0.15	0.81	0.09	0.91	0.91	16	3.9 (78%)
	1986 [d]	0.14	0.28	0.12	0.15	0.29	0.25	0.56	15	1.8 (36%)
Vashon	1985	2.13	1.17	5.61	1.92	0.01	.1.43	0.42	14	12.6 (252%)
	1986	0.11	0.25	0.04	0.17	0.43	0.27	0.97	12	2.3 (46%)
Seattle	1985	1.44	0.72	0.39	0.90	0.21	0.66	1.64	29	5.9 (118%)
	1986	0.34	0.41	0.18	0.29	0.35	0.37	1.19	19	3.2 (64%)
Duvall	1985	2.88	0.84	0.55	0.52	0.36	0.99	3.72	9	9.8 (196%)
	1986	0.08	0.12	0.10	0.12	0.46	0.13	0.90	11	1.9 (38%)

[a] Regions are termed: upwind of the smelter ("clean" background), immediately downwind of the smelter ("Vashon"), the Seattle urban area and a rural area downwind of Seattle ("Duvall"). [b] The upwind area (both years 1985 and 1986) was used as class 1 previously and the Vashon area for year 1 was used as class 2 previously ("smelter influenced") while the Seattle, Duvall areas for both years and the Vashon area for 1986 only were termed class 3 in the principal components classification. [c] 1985 was the year when the smelter was operating. [d] 1986 was the year when the smelter was closed.

"other influences" (As and Sb fractions in crust and sea salt are quite low). Other than 6, these influences can be interpreted to represent either: (a) other crust-like sources such as coal ash, wood, and cement dust, (b) real crustal material with less Al than the global average, or (c) ambient samples that are mixtures of crust, wood, coal, and/or cement dust as well as previously deposited and resuspended As and Sb (from 90 years of smelter operation). For the purposes of the present analysis it is not necessary to specifically identify each of these influences.

The "other influence" (No. 6) presents challenges in order to apportion smelter, crustal, and sea salt contributions to rainwater because it probably does contain "true" sea salt. Table 4 also presents the calculated enrichment in the rainwater samples of several trace elements relative to sea salt (calculated from Na); clearly Na, Mg, Br, and possibly K and Ca owe part of their composition to the pure sea salt "fingerprint" (No. 2). Thus the discriminant PLS training model was successively adjusted (until the guess matched the model reprediction of the sea salt content) to allow these rainwater samples comprising influence 6 to represent both sea salt and a new influence 6′ (that presented in Tables 3 and 4 but corrected for sea salt).

With the revised discriminant PLS model, as described above, seven source contributions (as mixing ratios) were calculated for each rainwater sample using 12 components. Reprediction of the mixing fractions for the "pure" source profiles using the PLS training model produced mixing factions within 5% of the expected values of 1.0 or 0.0 (for profile 6 these values were 0.52 sea salt and 0.48 source 6′); this confirmed that the seven influences could be independently predicted by the model (they were not colinear). As is expected for this type of a least squares technique, some individual predictions that were non-physical were obtained (> 1.0 or < 0). These were handled as follows: (a) three samples that were quite badly predicted (evidently outside of the mixing space defined by the hypercube but in a higher dimension than inspected previously) were discarded and (b) samples with predicted mixing ratios (for an individual "source") that were small and negative (> -0.1) were set to zero and the other mixing ratios for that sample were adjusted appropriately (mixing ratios sum to 1.0 for any sample). This last approach was previously used on a synthetic data set and produced quite accurate results [23].

Since the data were normalized (to the sample strength expressed as the sum: [Al + V + Mn +

TABLE 6

Crustal, sea salt, smelter and four other contributions to rainwater Al, V, Mn, As, Na, Ca, Mg and Fe concentrations (calculated from the PLS model)

[Values are (ng l^{-1}) contributions to the grand average "total" composition (125 samples). The chemical composition of the three profiles (1–3) and the four "other contributions" (4–7) derived from the data is presented in Table 3]

Source	Al	V	Mn	As	Na	Ca	Mg	Fe
Crust (1)	79.4	0.20	1.64	0.0	28	61	53	52
Sea salt (2)	0	0	0	0	989	46	223	0
Smelter (3)	0	0	0	2.16	0	0	0	< 1
ID 4	18.7	0.12	0.60	0.17	54	25	0	7
ID 5	6.4	0.04	0.13	0.02	51	5	0	4
ID 6	0.7	0.01	0.07	0	0	4	0	1
ID 7	21.6	0.10	1.27	0.03	36	40	0	30
Fraction due to: 1 + 2 + 3 (a priori profiles)	63%	45%	43%	90%	88%	59%	100%	53%
Fraction due to: 1 + 7 ("crust-like")	80%	68%	77%	1%	6%	56%	29%	83%

As + Na]; each element described as its sample value divided by the data grand mean) prior to PLS, the calculated mixing fractions were weighted by each sample's "strength" to determine useful source contributions to the rainwater chemical composition. Table 5 presents the results of this analysis; data are stratified into somewhat narrower geographical classes in order to further evaluate whether there were spatial tendencies in the predicted source contributions. Except for influences 5 and 7, all sources were calculated to have reduced contributions in year 2 than in year 1 with this change smallest in the upwind region.

Table 6 presents the PLS predictions of the seven source contributions to rainwater Al, V, Mn, As, Na, Ca, Mg, and Fe mass concentrations (as ng l^{-1}). The three a priori profiles account for ca. 90% of the As, Mg and Na, ca. 60% of the Al and Fe, and ca. 50% of the Mn, Ca, and V (sum of all individual sample predictions compared to the data grand mean) concentrations.

DISCUSSION

In previous analyses of data from the same experiment, ANOVA confirmed a significant smelter effect (upwind–downwind difference that disappeared when the smelter closed, termed a year-class interaction in ANOVA) on precipitation pH and sulfate for a region (class 2 above) extending out to 25 km northeast of the smelter [6]. Furthermore, the upwind sampling sites received cleaner rain in the year after the smelter closed than during the first year. The reasons for the lack of a detectible smelter contribution further than 25 km and the change in the background concentrations from year 1 to year 2 proved to be somewhat of a puzzle until the analysis presented herein.

The initial PCA, although accounting for 63% of the data variance, did not provide answers to the above questions. It was expected that the "first" PCA performed on all samples might not be sensitive enough to characterize some of the secondary (unknown) emission source influences on these rainwater samples. This situation would result whenever several "classes" of data are analyzed together because of blurring of source signatures important for portions of the sampling network by conflicting signatures for samples from areas affected by different sources. For example, if Fe was correlated to As and Sb near the smelter but to entirely different elements elsewhere, the net result of a PCA on all samples taken together could be only a weak correlation between Fe and the other relevant elements. Thus it was necessary to stratify the chemical data into regions of fairly homogeneous emission source influences.

Once the data were stratified, clear spatial differences were observed; the rainwater collected in the region immediately downwind of the smelter had significantly different ($p < 0.01$) trace element concentrations than did the upwind or further downwind areas. The PLS model allowed three obvious source contributions to be accounted for prior to the resolution of other source influences and PLS predicted actual mass concentrations. Atmospheric chemists typically find mass contributions much more tangible and relevant than how much variance was accounted for by a model (as in PCA).

As expected, the smelter impact is pronounced for the area ("Vashon") immediately downwind of the copper smelter and nearly vanishes during year 2 (smelter closed). This 'smelter influence' on rainwater composition generally decreases with downwind distance and extends out to 60 km downwind into the "Seattle" and "Duvall" areas (previously referred to as class 3). The previous analysis of sulfate and hydrogen ion data (using ANOVA; Ref. 8) failed to detect a smelter influence on this class 3 region, presumeably due to the influence of other sulfur sources in the area.

The spatial patterns in the contributions of the "pure" sea salt and crustal sources are not as coherent as is the smelter influence. As expected by their proximity to the Puget Sound waters, the "upwind" (class 1) and Vashon areas had the largest sea salt contributions. The decrease in the sea salt contribution from year 1 to year 2 in all areas illustrates one problem inherent in the data set. Despite the fact that meteorology was carefully tracked prior to and after sampling, the rain

events for the two years apparently differ in more ways than just the smelter's contribution (sea salt and crustal contributions decreased as did "other influences" 4 and 6).

The interpretation of the crustal source contribution is complicated by the likelihood that the "other influences" (4–7) may, in reality, be largely crustal material with either changes in elemental composition (e.g., basalt vs. granite, coal ash or cement) or errors in the chemical analysis (e.g., low Al). Further insight is obtained by prediction of individual trace element concentrations and inspection of the contributions of crust (1) and influences 4, 5, 6, and 7 to eight species of particular interest.

If "other influence" 7 is considered to represent crustal material with different elemental ratios (especially high Sm and La) than average, then the two crust-like (1 and 7) sources contribute 56% to 83% of the Al, V, Mn, Fe, and Ca concentrations but only 1% of the As, 6% of the Na, and 29% of the Mg concentrations. Influence 4 does contribute 7% the As, Sb (not shown), and Fe concentrations and displays temporal and spatial patterns (Table 5) that are similar to those of the smelter "signature" itself. This influence could reasonably be attributed to smelter particulate emissions that might logically (copper ore is Cu, As, Sb, Se, and Pb enriched crustal material) have similarities to crustal sources (Mn, Ti, Ca, Fe crustal enrichment factors < 3 for influence 4).

Influence 5 appears to be sea salt related but with more Al, Dy, and As than sea salt. Influence 6, already divided into sea salt and 6′ (non-sea salt portion of source 6) is enriched in Dy, Ca, Mn, Ti, K, and Fe relative to crustal material. Table 5 reveals that "other influence" 6 decreases in importance as one travels to the northeast (i.e., from Vashon to Seattle to Duvall). While it is somewhat tempting to identify 6 as the particulate emissions from a coal-fired power plant that is located about 100 km south of the copper smelter, inspection of the predictions for individual storms does not quite reveal the coherent spatial pattern (a "plume") that could confirm this speculation. In contrast, true crustal material might exhibit a somewhat patchy spatial and temporal pattern because of variation in its generation (wind speed), transport (gravitational settling), and, in particular, its limited ability to serve as a cloud condensation nuclei (probably the dominant source for incorporating many of these 17 trace elements into precipitation) [30]. Thus, influences 4 and 5 have not been lumped into the summaries of sea salt and crustal contributions to individual trace elements (Table 6). It is important to note that any interpretation of influences 4 to 7 probably would have failed without an experimental design that allowed spatial and temporal comparisons.

Conclusions

The contributions of sea salt, crustal material, and a copper smelter to rainwater trace element concentrations were calculated along with four other influences. The four other factors were evaluated by examining their chemical composition and by taking advantage of an experimental design that allowed spatial and temporal comparisons. The areal extent of the smelter's contribution to trace element concentrations in rain was seen out to the 60 km extent of the sampling network; this distance is greater than was revealed for smelter influences on sulfate and pH in a previous analysis. Besides changes associated with the smelter's closure, additional differences between the two years' rain event were identified and quantified.

Rich Peterson is thanked for performing the INAA analyses and Tim Larson for many stimulating conversations on all phases of this project. The present analysis was supported by NSF grant ATM-9118316 (Atmospheric Chemistry Desk). The experimental work was previously supported by the USEPA (EPA-DWA 89930059) via Battelle PNL (USDOE #DE-AC06-76RLO-1830).

REFERENCES

1 P.K. Hopke, E.S. Gladney, G.E. Gordon, W.H. Zoller and A.G. Jones, Atmos. Environ., 10 (1976) 1015.
2 G.E. Gordon, Environ. Sci. Technol., 22 (1988) 1132.
3 R.J. Vong, B.H. Bailey, M.J. Markus and V.A. Mohnen, Tellus, 42B (1990) 345.

4 D. Savoie, Atmos. Environ, 22 (1988) 1957.
5 R.J. Vong, H.-C. Hansson, H.R. Ross, D.S. Covert and R.J. Charlson, J. Geophys. Res., 93 (D2) (1988) 1625.
6 R.J. Vong, L. Moseholm, D. Covert, P. Sampson, J. O'Loughlin, M. Stevenson, R. Charlson, W. Zoller and T. Larson, J. Geophys. Res., 93 (D6) (1988) 7169.
7 R.J. Vong, T. Larson and W. Zoller, Chemom. Intell. Lab. Syst., 3 (1988) 99.
8 R.J. Vong, T.V. Larson, W.H. Zoller, D.S. Covert, R.J. Charlson, I. Sweet, R. Peterson, T. Miller, J. O'Loughlin and M. Stevenson, in R.W. Johnson and G.E. Gordon (Eds.), The Chemistry of Acid Rain: Sources and Atmospheric Processes (ACS Symp. Ser. 349), American Chemical Society, Washington, DC, 1987, p. 204.
9 R.E. Peterson, PhD thesis, University of Washington, Seattle, WA, 1991.
10 M.M. Minor and S.R. Garcia, in O.K. Harling, L. Clark and P. Von der Hardt (Eds.), Use and Development of Low and Medium Flux Research Reactors, Proc. Int. Symp., Massachusetts Institute of Technology, Cambridge, MA, 1983.
11 L.A. Currie, Anal. Chem., 40 (1968) 586.
12 H.-C. Hansson, B.G. Martinsson and H.O. Lannefors, Nucleic Instrum. Methods, B3 (1984) 483.
13 E. Johansson, S. Wold and K. Sjodin, Anal. Chem., 56 (1984) 1685.
14 E. Swietlicki, PhD thesis, Lund University, Lund, Sweden, 1989.
15 K.J. Noone, J.A. Ogren, A. Hallberg, H.-C. Hansson, A. Wiedensohler and E. Swietlicki, Tellus, 44B (1993) 581.
16 L.J. Ko, M.S. Thesis, Oregon State University, Corvallis, OR, 1992.
17 D.J. Albert and P.K. Hopke, Atmos. Environ., 14 (1980) 1137.
18 P.K. Hopke, Receptor Modeling in Environmental Chemistry, Wiley-Interscience, New York, 1985.
19 S. Wold, Pattern Recognition, 8 (1976) 127.
20 P. Geladi and B.R. Kowalski, Anal. Chim. Acta, 185 (1986) 1.
21 S. Wold, Technometrics, 20 (1978) 397.
22 H. Wold, in F.N. David (Ed), Festschrift for J. Neyman, Wiley, New York, 1966, p. 411.
23 R.J. Vong, P. Geladi, S. Wold and K. Esbensen, J. Chemom., 2 (1988) 281.
24 H.D. Holland, Chemistry of the Atmosphere and Oceans, Wiley, New York, 1978.
25 S.R. Taylor, Geochim. Cosmochim. Acta, 28 (1964) 1273.
26 A. Akulsweeny, R. Vong and R. Peterson, unpublished aircraft data, University of Washington, Seattle, WA, 1986.
27 J.M. Wilder, T. Tucker and R. Welch, Characterization of Visible Plumes from a Copper Smelter, presented at Pac. N.W. Intl. Section-Air Poll. Contr. Assoc. Annual Meeting, Vancouver, November 17, 1982.
28 L. Moseholm, Environ. Pollut., 53 (1988) 313.
29 S. Wold, C. Albano, G. Bloomquist et al., in A. Hoskuldsson, K. Konradson, B.S. Jensen and K. Esbensen (Eds.), Proceedings of the Symposium on Applied Statistics, NEUCC RECAU RECKU, Copenhagen, 1981.
30 H.R. Pruppacher and J.D. Klett, Microphysics of Cloud and Precipitation, Reidel, Dordrecht, 1978.

Analytica Chimica Acta, 277 (1993) 405–419
Elsevier Science Publishers B.V., Amsterdam

Empirical Bayes estimation in factor analysis for aerosol mass apportionment

Leon Jay Gleser and Hefei Yang

Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260 (USA)

(Received 15th July 1992; revised manuscript received 10th February 1993)

Abstract

In virtually all applications of factor models to aerosol mass apportionment, two important implicit assumptions are made: (a) that the source composition profile matrices remain constant over samples, and (b) that the (diagonal) error covariance matrices are the same from sample to sample. As observed by Gleser [Chemom. Intell. Lab. Syst., 10 (1991) 45] and Currie [Chemom. Intell. Lab. Syst., 10 (1991)], neither of these assumptions is generally true. In this paper, a factor analysis model is presented that accounts for variation over samples of both the source composition profiles and the error covariance matrices. Based on this model, point estimators and confidence set estimators are obtained for the vectors of source contributions using an empirical Bayes approach.

Keywords: Aerosols; Bayesian estimation; Factor analysis

Several broad classes of mathematical models have been used to apportion the aerosol measured at a receptor site to its likely sources. Of these, the chemical mass balance (CMB) model is most frequently used. This model is based on the assumption of proportional mass conservation as particles travel from source(s) to the receptor; that is, even if particles are deposited or dispersed in transit, no change is assumed to occur in the *relative amounts* of the chemical species (elements). Thus, mathematically, the measurements C_{it} of concentrations of aerosol property i at a receptor at time t and the true concentrations c_{it} obey the model:

$$C_{it} = c_{it} + \epsilon_{it}$$

$$c_{it} = \sum_{j=1}^{r} a_{ij}(t) s_{jt}, \quad i = 1, \cdots, m \qquad (1)$$

for $t = 1, \cdots, T$, where the ϵ_{it} are random measurement errors, s_{jt} is the mass contribution from the jth source at time t, and $a_{ij}(t)$ is the mass fraction of aerosol property i from source j at time t. The matrix $\mathbf{A}(t) = [a_{ij}(t)]$ is the mass fraction or source composition profile (SCP) matrix, and $\mathbf{s}(t) = (s_{1t}, s_{2t}, \cdots, s_{rt})'$ is the vector of source contributions.

Letting $\mathbf{C}(t) = (C_{1t}, C_{2t}, \cdots, C_{mt})'$, and $\boldsymbol{\epsilon}(\mathbf{t})$ be vector of the ϵ_{it}, defined in a similar fashion, the model 1 in vector-matrix form becomes:

$$\mathbf{C}(t) = \mathbf{A}(t)\mathbf{s}(t) + \boldsymbol{\epsilon}(t)$$

$$t = 1, \cdots, T \qquad (2)$$

In the chemical mass balance model, it is assumed that the SCP matrices $\mathbf{A}(t)$ have the same value A for all times or samples t, and that the random error vectors $\boldsymbol{\epsilon}(t)$ have zero mean vector and a constant covariance matrix $\mathbf{B}$, which is assumed to be diagonal. Thus, the chemical mass balance model takes the form

$$\mathbf{C}(t) = \mathbf{A}\mathbf{s}(t) + \boldsymbol{\epsilon}(t)$$

$$t = 1, \cdots, T \qquad (3)$$

Correspondence to: L.J. Gleser, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260 (USA).

which is recognizable as a classical factor analysis model (when $T > 1$). For $T = 1$, the chemistry literature refers to the model 3 as the chemical mass balance (CMB) model, whereas when $T > 1$, so that the aerosol property samples are taken in more than one size range or at more than one time, the model 3 is called the multivariate analysis model (MA). One important distinction between the CMB and MA models is that in the CMB model there is no approach available for solving 3 for the vector $\mathbf{s} = \mathbf{s}(1)$ of source contributions without prior knowledge of the SCP matrix **A**, and also the number r of potential sources [1,2].

Applications of the CMB model use weighted regression, effective variance calculations or ridge regression to estimate source contributions (and any other unknown parameters). Effective variance calculations for the CMB model have the desirable property that aerosol properties with higher precision in both the source measurements (of **A**) and receptor measurements are given greater influence in the effective variance least squares solution than are aerosol properties with lower precision [3,4], Ridge regression is applied to cope with the "multicollinearity" problem, which usually occurs in environmental studies when two sources have very similar chemical composition or when one source composition is nearly identical to a linear combination of other source compositions. In general, the successful applications of these methods to CMB depend heavily on reliable estimates of the SCP matrix **A**, and the knowledge of the number r of emission sources. In addition, when the multicollinearity phenomenon exists among source composition profiles, the source apportionment abilities of these methods can be severely hindered [5].

Various factor analysis estimation methods have been developed for the MA model. These methods are intended to (1) identify the number of major active sources, (2) estimate the SCP matrix **A** for such sources, and (3) estimate the vector $\mathbf{s}(t)$ of true source contributions in each collected sample, $t = 1, \cdots, T$. The two most popular such methods are principal component factor analysis (PCFA) and target transformation factor analysis (TTFA) [2,6]. Because the parameters (such as **A**) of a factor analysis model such as (3) are not identifiable by the data without identifying restrictions, each of these methods takes a different approach to letting the data suggest the restrictions to be used. Centering the original data C_{it} about either their means over the index t (PCFA) or about the origin (TTFA), a principal components (eigenvalue–eigenvector) analysis of the matrix of cross products is performed, so that a set of basis vectors for the initial representation of the source contributions can be obtained. If there were errors present and if the contributing sources were mutually independent, the number r of active sources would be in theory equal to the number of positive eigenvalues obtained. Instead, a judgement must be made of which of the smallest eigenvalues is only reflecting measurement error; deleting these yields a value for the number r of active sources. Several methods have been suggested for making such decisions [7,8]. Once r has been determined, interest then centers on determining **A**. In the case of PCFA, an estimate of **A** is obtained by assuming that it has "simple structure" [9] and rotating the eigenvectors corresponding to the r largest eigenvalues so as to achieve such a structure using one of several orthogonal rotation algorithms currently available. The VARIMAX rotation algorithm is one of the most applicable ones [10].

In contrast, the TTFA approach starts by assuming that the columns of **A** should be close to one or more test vectors that reflect the source contribution profiles of various types of sources. The eigenvectors corresponding to the largest r eigenvalues are rotated through an iterative refinement process relative to such test vectors. Because the test vectors are not necessarily in the same scale as the eigenvectors, the columns of the SCP matrix obtained by TTFA may not be comparable in scale to the true columns of **A**. The scaling factors needed to modify **A** to be scale-compatible are found by using a multiple regression of measured mass from each source on the source contribution columns determined by TTFA.

Once the matrix **A** has been determined, CMB methods can be used to estimate the source contribution vectors $\mathbf{s}(1), \cdots, \mathbf{s}(T)$. It is also desirable

to provide measures of precision for these estimates, and perhaps confidence intervals for the elements of the source contribution vectors. Here, there is considerable confusion in the literature. One measure of precision that is often given comes from assuming that the matrix **A** is known and equal to the value given for it by the PCFA or TTFA fitting method. In this case, the covariance matrix of an estimator of $\mathbf{s}(t)$ can be obtained from the theory of generalized least squares [11, Section 14.7]. However, this approach ignores the errors involved in estimating the matrix **A**. (In defense of this omission, such errors of estimation can be assumed to be small relative to the error involved in fitting $\mathbf{s}(t)$ given **A** when the number T of samples is large.)

Investigators who take account of the errors involved in fitting **A** usually use the large-sample (large T) formulas provided in the factor analysis literature. These formulas assume that the model has been identified in a pre-specified way, rather than having the identifying restrictions determined by the data (as in the VARIMAX or TTFA approaches). Consequently, such formulas are not, in the strict sense, applicable. Further, the large-sample approximations used do not necessarily provide good approximations for samples of moderate size T; this is particularly the case for the large-sample approximate confidence intervals given in the statistical literature [12].

Both of these approaches to providing measures of precision for the estimation of the vector $\mathbf{s}(t)$ of source contributions also overlook the possibility that something may be known about the variation of $\mathbf{s}(t)$ from sample to sample. If such information is available, in the form of a parameterized family of distributions for the $\mathbf{s}(t)$, improved point and interval estimators for $\mathbf{s}(t)$ can be obtained [11, Section 14.7], at least when one is willing to treat the fitted value of **A** as if it were the known true value.

As noted by Cheng and Hopke [13], it is not realistic to assume that the SCP matrices $\mathbf{A}(t)$ are constant over samples t. Consequently, estimates of the SCP matrix taken at a particular sample t_0 may not be valid for other samples $t \neq t_0$. In addition, the mass conservation assumptions that underly the CMB (and MA) model are never perfectly satisfied because there may be some mass loss due to precipitation, dispersal or chemical reaction along the path taken by particles to the receptor, and there is no reason a priori to assume that such losses are proportionately the same for all chemical species. Further, there may be contributions in some samples from sources not accounted for in the MA analysis because these unidentified sources were not present in all samples. Finally, in virtually all applications of MA models to aerosol mass apportionment, the error covariance matrices of measurement are assumed to be the same, and diagonal, from sample to sample. As Currie [14] points out, this is generally untrue because chemical measurement error increases with increasing concentration and the concentration of a given chemical species may vary widely over samples depending on both the relative and absolute amounts of the predominant components contributing to each sample.

What is needed, then, is a model that reflects variation of the SCP and error-covariance matrices over samples and also accounts for the possibility of intermittent source contributions form unidentified sources. To a statistician, an attractive way to model such variability is to assume that the SCP and error-covariance matrices are random matrices with a known parameterized family of distributions. To reflect the variation of the source contribution vectors $\mathbf{s}(t)$, $t = 1, \cdots, T$, one might also allow these vectors to be random, with a known parameterized family of distributions. Because chemists tend to think of such vectors and matrices as being fixed parameters, rather than random quantities, such models are not necessarily frequentist (empirical) in nature, but rather are a way of modeling an investigator's belief. The modeling of belief about parameters in a probabilistic fashion is the domain of the Bayesian approach to inference; if, however, these parameters actually do vary over samples in a fashion that resembles the variation of random quantities (e.g., their empirical distribution functions approach a limit as the number T of samples grows large), then treating such parameters as random quantities has an empirical justification, and methods following from such an as-

sumption of randomness are called empirical Bayes methods.

Below, a factor analysis model for aerosol mass apportionment is presented that permits SCP matrices to vary from sample to sample, and also accounts for heterogeneous covariance matrices for the disturbances in the CMB model (which now are allowed to reflect the combination of measurement error, loss of mass in transit, and intermittent contributions from unidentifiable sources). An empirical Bayes method for fitting this model is given also and illustrated with simulated data based on actual aerosol data reported by Albert and Hopke [15].

FACTOR ANALYSIS MODEL

The model proposed in this paper has the form as presented in Eqn. 1, or Eqn. 2 in vector–matrix form, but now the quantities ϵ_{it} represent the sum of: (i) errors of measurement at the receptor, (ii) mass from identified sources lost in transit to the receptor and (iii) mass added at the receptor from unidentified sources. It is assumed for each chemical species i that, on the average over samples t, the mass lost and the mass added balance out, so that the mean of ϵ_{it} is 0. On the other hand, the mass additions and losses can be correlated across chemical species i. Thus, the covariance matrix $\mathbf{V}(t)$ of $\epsilon(t) = (\epsilon_{1t}, \cdots, \epsilon_{mt})'$ can be any positive definite matrix. Because the processes providing the disturbance vector $\epsilon(t)$ may vary over samples t, the covariance matrices $\mathbf{V}(1)$, $\mathbf{V}(2), \cdots, \mathbf{V}(T)$ of the disturbance vectors are not assumed to be related to one another in any functional way. Following tradition in factor analysis, it is assumed that the disturbance vectors $\epsilon(1), \cdots, \epsilon(T)$ have multivariate normal distributions:

$$\epsilon(t) \sim MVN[\mathbf{0}, \mathbf{V}(t)], \quad t = 1, \cdots, T \tag{4}$$

The variation of the SCP matrices $\mathbf{A}(t)$ and the disturbance covariance matrices $\mathbf{V}(t)$ over samples is modeled by a joint probability distribution for the elements of these two matrices. This distribution is constructed as follows.

First, conditional on the elements $v_{gh}(t)$ of $\mathbf{V}(t)$, the elements $a_{ij}(t)$ of $\mathbf{A}(t)$ have a multivariate normal distribution, where $a_{ij}(t)$ has mean λ_{ij} and the covariance between $a_{ij}(t)$ and $a_{kl}(t)$ is

$$\operatorname{cov}[a_{ij}(t), a_{kl}(t)] = h_{ik} v_{jl}(t)$$

$$1 \le i, k \le r, \ 1 \le j, l \le m$$

Here,

$$\boldsymbol{\Lambda} = [(\lambda_{ij})], \ \mathbf{H} = [(h_{ij})]$$

are $m \times r$ and $r \times r$ matrices of constants, respectively. The probability density function of $\mathbf{A}(t)$ given $\mathbf{V}(t)$ thus has the form:

$$f[\mathbf{A}(t) \mid \mathbf{V}(t), \boldsymbol{\Lambda}, \mathbf{H}]$$

$$= \{\exp\{-(1/2)\operatorname{tr}[(\mathbf{A}(t) - \boldsymbol{\Lambda})\mathbf{H}^{-1}(\mathbf{A}(t) - \boldsymbol{\Lambda})' \mathbf{V}^{-1}(t)]\}\}$$

$$\times \left(|\mathbf{V}(t)|^{r/2} |\mathbf{H}^{-1}|^{m/2} (2\pi)^{mr/2}\right)^{-1} \tag{5}$$

over all $m \times r$ matrices.

Note that the matrices $\boldsymbol{\Lambda}$ and $\mathbf{H}$ serve to parameterize the distribution 5. In applications, one would expect elements $a_{ij}(t)$, $a_{kl}(t)$ in different columns of $\mathbf{A}(t)$ to be unrelated, in terms of their variation over samples t, because the columns of $\mathbf{A}(t)$ give the contribution profiles of different sources $i \neq k$ and such profiles are determined almost entirely by chemical processes within each such source. In this case, $h_{ik} = 0$ for $i \neq k$, and $\mathbf{H}$ can be assumed to be a diagonal matrix. In general, the matrix $\mathbf{H}$ relates the variation in the columns of $\mathbf{A}(t)$ over samples t to the variation in the disturbance vectors $\epsilon(t)$. For example, small diagonal values of $\mathbf{H}$ would indicate that the sample-to-sample variation of the columns of $\mathbf{A}(t)$ is small relative to the sample-to-sample variation of the disturbances $\epsilon(t)$.

Second, the matrix $\mathbf{V}(t)$ is assumed to have an inverted Wishart distribution; that is, the inverse matrix $\mathbf{V}^{-1}(t)$ has a Wishart distribution with ν degrees of freedom and expected value $\nu\mathbf{B}^{-1}$. The density function of $\mathbf{V}(t)$ is thus

$$f[\mathbf{V}(t) \mid \nu, \mathbf{B}]$$

$$= \frac{K |\mathbf{B}|^{\frac{\nu}{2}} \exp\{-(1/2) \operatorname{tr}[\mathbf{B}^{-1}\mathbf{V}^{-1}(t)]\}}{|\mathbf{V}(t)|^{(\nu+m+1)/2}} \tag{6}$$

over the space of $m \times m$ positive definite matrices, where

$$K = \left[\prod_{l=1}^{m} \Gamma\left(\frac{\nu - l + 1}{2} \right) 2^{\frac{\nu m}{2}} \pi^{m(m-1)/2} \right]^{-1}$$

In this case, the mean of $\mathbf{V}(t)$ is equal to $(\nu - m - 2)^{-1}\mathbf{B}$, provided that $\nu > m + 2$.

Finally, $[\epsilon(t), \mathbf{V}(t), \mathbf{A}(t)]$, $t = 1, \cdots, T$, are assumed to be mutually independent. That is, the disturbance, random disturbance covariance matrix and SCP matrix for any sample t^* is statistically independently of the disturbances, random disturbance covariance matrices and SCP matrices for all other samples $t \neq t^*$. As at least a first approximation, this assumption appears reasonable.

In order to fit even the classical MA model (3), the source contribution vectors $\mathbf{s}(t)$ must vary over samples t. It will be assumed that the $\mathbf{s}(t)$ are random vectors having a common distribution with mean vector μ and covariance matrix $\mathbf{D}$, but the exact form of this common distribution need not be specified. It is also assumed that $\mathbf{s}(1), \cdots, \mathbf{s}(T)$ are mutually statistically independent, and that $\mathbf{s}(t)$ is independent of $[\epsilon(t), \mathbf{V}(t), \mathbf{A}(t)]$, $t = 1, \cdots, T$.

A classical confidence region for s(t)

Using Eqns. 2 and 4–6, it can be shown that for each sample t the conditional distribution of the observation $C(t)$ given the source contribution vector $\mathbf{s}(t)$ is a multivariate t-distribution [16] with density

$$\begin{aligned} f[\mathbf{C}(t) \mid \mathbf{s}(t)] &= \left(\left[\Gamma\left(\frac{\nu}{2} \right) / \pi^{m/2} \Gamma\left(\frac{\nu - m}{2} \right) \right] \right. \\ &\quad \times [1 + \mathbf{s}'(t)\mathbf{H}\mathbf{s}(t)]^{(\nu-m)/2} |\mathbf{B}|^{1/2} \Big) \\ &\quad \times [1 + \mathbf{s}'(t)\mathbf{H}\mathbf{s}(t) + (\mathbf{C}(t) - \boldsymbol{\Lambda}\mathbf{s}(t))' \\ &\quad \times \mathbf{B}(\mathbf{C}(t) - \boldsymbol{\Lambda}\mathbf{s}(t))]^{-\nu/2} \end{aligned}$$

over the space of all m-dimensional vectors $\mathbf{C}(t)$. From this result, it follows that $\mathbf{C}(t)$ given $\mathbf{s}(t)$ has the same distribution as

$$\boldsymbol{\Lambda}\mathbf{s}(t) + (1 + \mathbf{s}'(t)\mathbf{H}\mathbf{s}(t))^{1/2} \mathbf{B}^{-1/2} W^{-1/2} \mathbf{Z} \quad (7)$$

where

$$\mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{I}_m), \; W \sim \chi^2_{\nu - m}$$

$\mathbf{Z}$ and W are independent and $\mathbf{B}^{1/2}$ is the positive definite square root of $\mathbf{B}[\mathbf{B} = (\mathbf{B}^{1/2})(\mathbf{B}^{1/2})]$. Consequently,

$$\left(\frac{\nu - m}{m} \right) \frac{[\mathbf{C}(t) - \boldsymbol{\Lambda}\mathbf{s}(t)]' \mathbf{B} [\mathbf{C}(t) - \boldsymbol{\Lambda}\mathbf{s}(t)]}{1 + \mathbf{s}'(t)\mathbf{H}\mathbf{s}(t)} \sim F_{m,\nu-m} \quad (9)$$

where $F_{m,\nu-m}$ represents the F-distribution with m and $\nu - m$ degrees of freedom. It follows from Eqn. 9 that if $\boldsymbol{\Lambda}$, $\mathbf{B}$, $\mathbf{H}$ and ν are known,

$$\left\{ \mathbf{s}(t) : \left(\frac{m}{\nu - m} \right) \times \frac{[\mathbf{C}(t) - \boldsymbol{\Lambda}\mathbf{s}(t)]' \mathbf{B} [\mathbf{C}(t) - \boldsymbol{\Lambda}\mathbf{s}(t)]}{1 + \mathbf{s}'(t)\mathbf{H}\mathbf{s}(t)} \leq F_{m,\nu-m}(1 - \alpha) \right\} \quad (10)$$

where $F_{m,\nu-m}(1 - \alpha)$ is the $100(1 - \alpha)$th percentile of the $F_{m,\nu-m}$-distribution, provides a $100(1 - \alpha)\%$ confidence region for $\mathbf{s}(t)$. The advantage of the confidence region 10 is that is does not require knowledge of the distribution of $\mathbf{s}(t)$, or even the assumption that $\mathbf{s}(t)$ is random. One disadvantage of the region 10, however, is that it does not necessarily have a nice shape, so that it may not yield convenient regions (intervals) for the elements s_{it} of $\mathbf{s}(t)$. Indeed, letting $F^* = \nu - m/mF_{m,\nu-m}(1 - \alpha)$, the region 10 is equivalent to the region

$$[\mathbf{s}(t) - \hat{\mathbf{s}}(t)]' (\boldsymbol{\Lambda}'\mathbf{B}\boldsymbol{\Lambda} - F^*\mathbf{H}) [\mathbf{s}(t) - \hat{\mathbf{s}}(t)] \leq \mathbf{L}(t) \quad (11)$$

where

$$\hat{\mathbf{s}}(t) = (\boldsymbol{\Lambda}'\mathbf{B}\boldsymbol{\Lambda} - F^*\mathbf{H})^{-1} \boldsymbol{\Lambda}'\mathbf{B}\mathbf{C}(t), \quad (12)$$

$$L(t) = F^* - \mathbf{C}'(t)\mathbf{B}\mathbf{C}(t) + \mathbf{C}'(t) \times \left[\mathbf{B}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\mathbf{B}\boldsymbol{\Lambda} - F^*\mathbf{H})^{-1} \boldsymbol{\Lambda}'\mathbf{B} \right] \mathbf{C}(t)$$

assuming that $\mathbf{Q} = \boldsymbol{\Lambda}'\mathbf{B}\boldsymbol{\Lambda} - F^*\mathbf{H}$ is invertible. For the region 11 to be an ellipse, $L(t)$ must be greater than zero and $\mathbf{Q}$ must be a positive defi-

nite matrix. If $\mathbf{Q}$ is positive definite and $L(t) > 0$, then the following intervals provide simultaneous $100(1-\alpha)\%$ confidence intervals for the elements $s_{1t}, \cdots, s_{rt}$ of $\mathbf{s}(t)$:

$$\hat{s}_{it} \pm \sqrt{L(t)/q_{ii}}\,, \quad i = 1, \cdots, r \tag{13}$$

where $\hat{\mathbf{s}}(t) = (\hat{s}_{1t}, \cdots, \hat{s}_{rt})'$ and $q_{11}, \cdots, q_{rr}$ are the diagonal elements of $\mathbf{Q}$.

Another disadvantage of the region 10, and the associated simultaneous confidence intervals 13, is that the values of the parameters $\boldsymbol{\Lambda}$, $\mathbf{H}$, $\mathbf{B}$, and ν must be provided by the user. It is possible, however, to use the data to estimate some of these parameters, as shown in the next section.

EMPIRICAL BAYES APPROACH

Provided that a user is able to specify the value of the matrix $\mathbf{H}$, it is possible to use the data $\mathbf{C}(1), \cdots, \mathbf{C}(T)$ to estimate values of $\boldsymbol{\Lambda}$, $\mathbf{B}$ and ν to use in Eqn. 13. The data can also provide estimates of the mean vector μ and covariance matrix $\mathbf{D}$ of the assumed common distribution of the source contribution vectors $\mathbf{s}(t)$. Such an approach is in the spirit of the "empirical Bayes" approach to inference treated in the statistical literature [17]. In such an approach, either maximum likelihood or method of moments estimation is typically used. The latter approach is preferable here, because the maximum likelihood approach requires us to specify the exact form of the common distribution of $\mathbf{s}(1), \cdots, \mathbf{s}(T)$ in order to obtain the marginal distribution of the data $\mathbf{C}(1), \cdots, \mathbf{C}(T)$, and also because this marginal distribution is mathematically complicated.

A method of moments approach makes use of familiar factor analysis methods. Estimators produced by such an approach are consistent; that is, the estimates have increasingly high probability of being arbitrarily close to the true values of the parameters being estimated as the number, T, of samples observed increases to infinity. Thus, when T is fairly large, such estimates can be treated as if they were the known, correct values of the parameters.

Let

$$\bar{\mathbf{C}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{C}(t),$$

$$\mathbf{U} = \frac{1}{T}\sum_{t=1}^{T} \left[\mathbf{C}(t) - \bar{\mathbf{C}}\right]\left[\mathbf{C}(t) - \bar{\mathbf{C}}\right]' \tag{14}$$

and let

$$\boldsymbol{\Phi} = \mathbf{D} + \mu\mu', \quad \boldsymbol{\Psi} = \left[\frac{1 + \mathrm{tr}\mathbf{H}\boldsymbol{\Phi}}{\nu - m - 2}\right]\mathbf{B}^{-1}$$

Note that $\boldsymbol{\Psi}$ is a diagonal matrix if and only if $\mathbf{B}$ is a diagonal matrix. To motivate the use of factor analysis to estimate the parameters, it can be shown from Eqn. 7 that

$$E\left[\mathbf{C}(t)\mathbf{C}'(t)\right] = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{15}$$

has the familiar factor analysis form, assuming that $\boldsymbol{\Psi}$ is diagonal. Thus, if $\mathbf{B}$ (and thus $\boldsymbol{\Psi}$) is diagonal, a standard factor analysis method applied to the matrix $\mathbf{U} + T\bar{\mathbf{C}}\bar{\mathbf{C}}'$ can be used to obtain an estimator $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$, and provide a preliminary estimator $\tilde{\boldsymbol{\Lambda}}$ of the SCP matrix $\boldsymbol{\Lambda}$. It is recommended that this first factor analysis step make use of the identifying conditions: $\boldsymbol{\Phi} = \mathbf{I}_r$, $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is a diagonal matrix [11,18]. The TTFA approach [2,6] can then be used to transform $\tilde{\boldsymbol{\Lambda}}$ to $\hat{\boldsymbol{\Lambda}} = \tilde{\boldsymbol{\Lambda}}\mathbf{G}$. This yields the estimator $\hat{\boldsymbol{\Lambda}}$ for $\boldsymbol{\Lambda}$, and also the estimator $\hat{\boldsymbol{\Phi}} = (\mathbf{G}'\mathbf{G})^{-1}$ for $\boldsymbol{\Phi}$. Note that an estimator of the mean vector μ of the distribution of the source contribution vectors $\mathbf{s}(t)$ is found as part of the TTFA approach: namely, $\hat{\mu} = (\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Lambda}})^{-1}\hat{\boldsymbol{\Lambda}}'\bar{\mathbf{C}}$. Thus, an estimator of the covariance matrix $\mathbf{D}$ of $\mathbf{s}(t)$ is

$$\hat{\mathbf{D}} = \hat{\boldsymbol{\Phi}} - \hat{\mu}\hat{\mu}'$$

provided that $\hat{\mathbf{D}}$ is positive semi-definite.

Let

$$x(t) = \mathbf{C}'(t)\left[\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\right] \times \mathbf{C}(t)$$

This can be obtained as the residual sum of squares obtained by fitting the model

$$\mathbf{C}(t) = \boldsymbol{\Lambda}\mathbf{s}(t) + \text{error}, \quad \text{cov(error)} = \boldsymbol{\Psi}$$

for $\mathbf{s}(t)$ using weighted least squares. Using Eqn. 7, it can be shown that, conditional on $\mathbf{s}(t)$,

$$x(t) \sim \left[\frac{1+\mathrm{tr}(\mathbf{H}\boldsymbol{\Phi})}{\nu-m-2}\right] \times \left\{\frac{m-r}{(\nu-m)[1+\mathbf{s}'(t)\mathbf{H}\mathbf{s}(t)]}\right\} F_{m-r,\nu-m}$$

from which it follows that

$$\frac{E[x^2(t)]}{E^2[x(t)]} = \left(\frac{m-r+2}{m-r}\right)\left(\frac{\nu-m-2}{\nu-m-4}\right),$$

$\nu \geq m+4$.

To make use of this fact, let

$$\hat{x}(t) = \mathbf{C}'(t)\left[\hat{\boldsymbol{\Psi}}^{-1} - \hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}\right)^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\right] \times \mathbf{C}(t)$$

and

$$Y = \frac{\frac{1}{T}\sum_{t=1}^{T}[\hat{x}(t)]^2}{\left[\frac{1}{T}\sum_{t=1}^{T}\hat{x}(t)\right]^2}$$

TABLE 1

Refined source profiles (mg g^{-1}). RAPS station 112. July and August, 1976. Samples from July 4th and 5th excluded. Fine fraction

Elements	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
Al	5.0	1.1	53.0	0.0	0.0
Si	0.0	1.9	130.0	0.0	7.0
S	0.02	240.0	19.0	6.0	0.0
Cl	2.4	1.1	0.0	4.6	22.0
K	1.4	1.6	15.0	5.7	48.0
Ca	11.0	0.0	16.0	34.0	1.2
Ti	0.0	0.7	2.5	110.0	0.0
Mn	0.0	0.0	0.7	4.8	8.6
Fe	0.0	1.1	36.0	90.0	36.0
Ni	0.08	0.04	0.042	0.011	0.7
Cu	0.6	0.01	0.0	0.0	8.7
Zn	0.8	0.0	0.0	3.7	65.0
Se	0.1	0.1	0.001	0.2	0.2
Br	30.0	0.03	2.5	0.0	0.05
Sr	0.09	0.01	0.15	0.1	0.005
Ba	0.7	0.05	0.07	28.0	0.5
Pb	107.0	6.5	5.0	0.0	46.0

Then an estimator of ν is

$$\hat{\nu} = m + 4 + \frac{2(m-r+2)}{(m-r)Y-(m-r+2)}$$

Finally, an estimator of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \frac{\left[1+\mathrm{tr}(\mathbf{H}\hat{\boldsymbol{\Phi}})\right]}{\hat{\nu}-m-2}\hat{\boldsymbol{\Psi}}^{-1}$$

The estimators $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Phi}}$, $\hat{\mathbf{B}}$ and $\hat{\nu}$ obtained above can be used in the confidence region 10 as approximations to the true values of these parameters. This usage overlooks the error in estimation of these quantities, but such error will be small relative to other sources of error or variation when T is large. Further refinements of the region 10 that take account of the fact that $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\mathbf{B}$ and ν are estimated are possible [19].

An empirical Bayes approach can also be used to give point estimates (actually predictions) of the source contribution vectors $\mathbf{s}(t)$, $t = 1, \cdots, T$. Although the conditional expected value of $\mathbf{s}(t)$ given $\mathbf{C}(t)$ would be the best predictor, calculation of the expected value requires specification of the distribution of $\mathbf{s}(t)$ and numerical evaluation of complicated integrals. Instead, the best linear [in $\mathbf{C}(t)$] approximation to this conditional expected value can be used (a "linear Bayes estimate"). This is

$$\tilde{\mathbf{s}}(t) = \mathbf{D}\boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\mathbf{D}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}[\mathbf{C}(t) - \boldsymbol{\Lambda}'\mu] + \mu$$

as can be seen from Eqns. 7 and 15 by noting that

$$\mathrm{cov}\begin{bmatrix}\mathbf{s}(t)\\ \mathbf{C}(t)\end{bmatrix} = \begin{bmatrix}\mathbf{D} & \mathbf{D}\boldsymbol{\Lambda}'\\ \boldsymbol{\Lambda}\mathbf{D} & \boldsymbol{\Lambda}\mathbf{D}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}\end{bmatrix}$$

Plugging in estimates of the parameters yields the "linear empirical Bayes" point estimate

$$\hat{\mathbf{s}}(t) = \hat{\mathbf{D}}\hat{\boldsymbol{\Lambda}}'\mathbf{U}^{-1}(\mathbf{C}(t) - \bar{\mathbf{C}}) + \hat{\mu} \qquad (16)$$

of $\mathbf{s}(t)$, $t = 1, \cdots, T$.

An example

Albert and Hopke [15] applied TTFA to various sets of aerosol data obtained during the Regional Air Pollution Study (RAPS) for St. Louis, Missouri. One data set which they used involved fine fraction particles obtained from samples collected at one receptor site (station 112) during

July and August, 1976 (samples from July 4th and 5th were excluded). Their analysis identified $r = 5$ sources: motor vehicle, sulfate, flyash/soil, paint and refuse. The number of chemical species measured was $m = 17$. The source profile matrix $\boldsymbol{\Lambda}$ estimated by Albert and Hopke for their data is given in Table 1. Using their summary results, a realistic choice of mean vector and covariance matrix for the source contribution vectors $\mathbf{s}(t)$ might be:

$$\boldsymbol{\mu} = (4.5, 19.0, 3.2, 0.41, 1.23)',$$

$$\mathbf{D} = \text{diagonal}(2.5^2, 12.9^2, 4.3^2, 1.0^2, 1.0^2)'$$

where diagonal $(d_1, d_2, \cdots, d_r)$ denotes an r-dimensional matrix having diagonal elements $d_1, d_2, \cdots, d_r$. Their results also suggest that the di-

TABLE 2

Samples of source contributions

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.18	2.79	0.18	0.98	1.24
2	8.49	9.33	5.93	0.06	4.06
3	3.20	45.33	11.87	1.50	1.41
4	7.78	14.16	14.22	1.93	1.06
5	4.72	48.92	12.20	4.63	4.77
6	4.11	7.43	6.97	0.80	1.51
7	12.21	23.81	5.56	0.47	0.17
8	4.77	26.07	1.12	0.55	1.27
9	8.61	37.62	2.53	0.90	1.92
10	6.17	32.89	7.63	0.01	3.58
11	7.95	21.30	13.86	2.43	4.52
12	1.69	3.82	4.78	3.18	3.13
13	4.30	24.38	9.92	2.32	0.48
14	11.13	5.75	1.85	5.11	1.59
15	3.12	37.71	9.01	1.35	3.65
16	5.09	25.22	13.06	0.45	3.03
17	0.63	9.17	6.90	0.14	4.81
18	6.20	19.93	1.28	0.01	1.56
19	3.64	8.61	10.33	1.31	2.21
20	8.73	39.47	4.79	0.74	3.74
21	9.00	49.66	5.34	1.12	0.42
22	5.18	29.61	6.48	2.78	2.62
23	1.47	41.41	3.98	1.51	0.80
24	1.56	33.51	6.86	0.69	1.07
25	7.44	44.75	5.92	1.55	3.35
26	11.87	31.15	5.74	1.64	0.49
27	7.28	1.28	3.08	2.23	0.43
28	6.04	12.78	5.73	0.82	2.26
29	7.79	15.77	3.36	1.01	0.23
30	14.95	34.89	3.47	3.28	3.15
31	4.16	31.77	6.02	1.58	2.30
32	2.87	77.93	0.98	2.01	6.39
33	4.34	15.44	2.36	1.22	1.50
34	9.19	38.77	7.55	2.68	2.91
35	8.19	7.45	8.17	3.02	1.58
36	0.63	42.90	8.69	3.97	1.50
37	7.52	20.66	6.23	1.16	2.07
38	8.81	12.93	15.46	1.83	3.93
39	10.56	17.85	9.72	0.62	1.72
40	7.52	6.26	9.16	1.87	3.00

TABLE 3

Point estimates (Eqn. 12) of source contributions ($\mathbf{H} = 10\mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.20	2.79	0.08	0.88	1.27
2	7.97	10.03	6.15	0.03	3.91
3	2.95	47.98	10.40	1.40	1.47
4	8.04	15.81	13.93	2.17	1.25
5	5.22	55.29	12.01	4.47	5.12
6	3.91	9.71	6.24	1.02	1.75
7	11.41	25.80	5.48	0.08	0.16
8	4.99	29.62	0.37	0.71	1.42
9	8.80	40.92	3.04	0.50	1.86
10	6.72	35.68	7.51	0.07	3.68
11	8.44	24.77	15.10	2.53	3.90
12	1.74	3.90	5.03	3.26	2.73
13	3.61	29.26	10.90	2.15	0.13
14	12.00	5.15	1.81	5.19	1.55
15	2.88	47.11	10.43	0.71	3.12
16	5.61	25.61	13.03	0.81	2.41
17	0.87	10.56	6.91	0.22	4.85
18	5.68	20.80	0.84	0.13	1.89
19	3.47	9.93	10.22	1.23	2.62
20	10.51	45.79	4.19	0.36	4.47
21	8.49	57.74	5.30	0.51	0.64
22	5.96	30.38	6.07	2.55	3.64
23	0.98	47.27	5.05	0.52	0.95
24	2.39	35.28	7.76	0.75	0.92
25	7.58	52.84	5.46	1.21	4.32
26	12.23	35.99	6.73	0.78	0.27
27	7.43	2.31	3.15	2.25	0.80
28	6.18	13.27	5.66	0.88	2.39
29	8.14	16.62	3.06	0.82	0.58
30	15.22	43.64	2.93	3.52	2.33
31	3.14	39.68	6.10	1.80	1.84
32	4.87	86.76	2.41	1.57	6.55
33	4.58	17.50	2.50	1.31	1.61
34	6.94	47.57	8.27	2.28	3.35
35	8.01	8.53	8.65	2.72	1.74
36	2.00	47.77	7.44	4.48	1.94
37	8.54	23.97	6.84	0.86	2.27
38	9.37	16.59	15.96	1.33	3.95
39	11.18	21.31	10.12	0.51	2.23
40	7.88	6.83	9.53	1.92	2.95

agonal elements of the sample standard deviation matrix of $\overline{\mathbf{C}}$ are

$$\boldsymbol{\Delta}^{1/2} = \text{diagonal}(24.0, 59.0, 32.0, 9.0, 9.0, 10.0, 13.0, 3.0, 19.0, 0.2, 2.0, 8.0, 0.2, 8.0, 0.1, 4.0, 53.0)'$$

where $(\boldsymbol{\Delta}^{1/2})^2 = \boldsymbol{\Delta}$.

Given the values of $\boldsymbol{\Lambda}$, μ, $\mathbf{D}$, by Eqn. 15, an estimate of $\boldsymbol{\Psi}$ is obtained as $\hat{\boldsymbol{\Psi}}$, a diagonal matrix with its diagonal elements equal to those of $\sqrt{T}(\boldsymbol{\Delta}) - \boldsymbol{\Lambda}(\mathbf{D} + \mu\mu')\boldsymbol{\Lambda}'$. Because Albert and Hopke [15] do not give the individual data vectors $\mathbf{C}(1), \cdots, \mathbf{C}(T)$, and because it is of interest to see how the methods given in this paper perform on data that fit the model presented earlier in this

TABLE 4

Simultaneous 95% confidence intervals (Eqn. 13) for the elements of the source contribution vectors ($\mathbf{H} = 10\mathbf{I}_5$).

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.20 ± 0.59	2.79 ± 1.60	0.08 ± 0.47	0.88 ± 0.25	1.27 ± 0.36
2	7.97 ± 1.88	10.03 ± 5.05	6.15 ± 1.50	0.03 ± 0.78	3.91 ± 1.12
3	2.95 ± 5.26	47.98 ± 14.16	10.40 ± 4.21	1.40 ± 2.19	1.47 ± 3.15
4	8.04 ± 2.49	15.81 ± 6.69	13.93 ± 1.99	2.17 ± 1.04	1.25 ± 1.49
5	5.22 ± 5.93	55.29 ± 15.96	12.01 ± 4.74	4.47 ± 2.47	5.12 ± 3.55
6	3.91 ± 1.26	9.71 ± 3.38	6.24 ± 1.00	1.02 ± 0.52	1.75 ± 0.75
7	11.41 ± 3.11	25.80 ± 8.38	5.48 ± 2.49	0.08 ± 1.30	0.16 ± 1.86
8	4.99 ± 3.71	29.62 ± 9.97	0.37 ± 2.96	0.71 ± 1.54	1.42 ± 2.22
9	8.80 ± 4.94	40.92 ± 13.28	3.04 ± 3.95	0.50 ± 2.06	1.86 ± 2.95
10	6.72 ± 4.80	35.68 ± 12.91	7.51 ± 3.84	0.07 ± 2.00	3.68 ± 2.87
11	8.44 ± 3.87	24.77 ± 10.41	15.10 ± 3.10	2.53 ± 1.61	3.90 ± 2.32
12	1.74 ± 1.00	3.90 ± 2.70	5.03 ± 0.80	3.26 ± 0.42	2.73 ± 0.60
13	3.61 ± 3.48	29.26 ± 9.35	10.90 ± 2.78	2.15 ± 1.45	0.13 ± 2.08
14	12.00 ± 1.63	5.15 ± 4.38	1.81 ± 1.30	5.19 ± 0.68	1.55 ± 0.97
15	2.88 ± 5.60	47.11 ± 15.05	10.43 ± 4.47	0.71 ± 2.33	3.12 ± 3.35
16	5.61 ± 3.24	25.61 ± 8.71	13.03 ± 2.59	0.81 ± 1.35	2.41 ± 1.94
17	0.87 ± 1.40	10.56 ± 3.76	6.91 ± 1.12	0.22 ± 0.58	4.85 ± 0.84
18	5.68 ± 1.20	20.80 ± 3.22	0.84 ± 0.96	0.13 ± 0.50	1.89 ± 0.72
19	3.47 ± 1.82	9.93 ± 4.91	10.22 ± 1.46	1.23 ± 0.76	2.62 ± 1.09
20	10.51 ± 5.72	45.79 ± 15.38	4.19 ± 4.57	0.36 ± 2.38	4.47 ± 3.42
21	8.49 ± 7.47	57.74 ± 20.11	5.30 ± 5.98	0.51 ± 3.11	0.64 ± 4.47
22	5.96 ± 1.64	30.38 ± 4.42	6.07 ± 1.31	2.55 ± 0.68	3.64 ± 0.98
23	0.98 ± 5.73	47.27 ± 15.42	5.05 ± 4.58	0.52 ± 2.39	0.95 ± 3.43
24	2.39 ± 3.66	35.28 ± 9.85	7.76 ± 2.93	0.75 ± 1.53	0.92 ± 2.19
25	7.58 ± 6.21	52.84 ± 16.70	5.46 ± 4.96	1.21 ± 2.59	4.32 ± 3.72
26	12.23 ± 4.63	35.99 ± 12.45	6.73 ± 3.70	0.78 ± 1.93	0.27 ± 2.77
27	7.43 ± 0.95	2.31 ± 2.56	3.15 ± 0.76	2.25 ± 0.40	0.80 ± 0.57
28	6.18 ± 1.91	13.27 ± 5.14	5.66 ± 1.53	0.88 ± 0.80	2.39 ± 1.14
29	8.14 ± 2.20	16.62 ± 5.93	3.06 ± 1.76	0.82 ± 0.92	0.58 ± 1.32
30	15.22 ± 5.35	43.64 ± 14.39	2.93 ± 4.28	3.52 ± 2.23	2.33 ± 3.20
31	3.14 ± 5.17	39.68 ± 13.90	6.10 ± 4.13	1.80 ± 2.15	1.84 ± 3.09
32	4.87 ± 6.76	86.76 ± 18.18	2.41 ± 5.40	1.57 ± 2.81	6.55 ± 4.04
33	4.58 ± 2.24	17.50 ± 6.02	2.50 ± 1.79	1.31 ± 0.93	1.61 ± 1.34
34	6.94 ± 5.65	47.57 ± 15.20	8.27 ± 4.52	2.28 ± 2.35	3.35 ± 3.38
35	8.01 ± 1.62	8.53 ± 4.35	8.65 ± 1.29	2.72 ± 0.67	1.74 ± 0.97
36	2.00 ± 5.72	47.77 ± 15.38	7.44 ± 4.57	4.48 ± 2.38	1.94 ± 3.42
37	8.54 ± 2.60	23.97 ± 7.00	6.84 ± 2.08	0.86 ± 1.08	2.27 ± 1.56
38	9.37 ± 3.19	16.59 ± 8.57	15.96 ± 2.55	1.33 ± 1.33	3.95 ± 1.91
39	11.18 ± 3.24	21.31 ± 8.71	10.12 ± 2.59	0.51 ± 1.35	2.23 ± 1.94
40	7.88 ± 1.85	6.83 ± 4.96	9.53 ± 1.48	1.92 ± 0.77	2.95 ± 1.10

paper, the values of Λ, μ, $\mathbf{D}$ and $\Psi^{1/2}$ were used to simulate 40 samples of source contribution vectors $\mathbf{s}(1), \cdots, \mathbf{s}(40)$ and measured concentration vectors $\mathbf{C}(1), \cdots, \mathbf{C}(40)$. Rather than simulate $\mathbf{C}(t)$ by first simulating $\mathbf{A}(t)$ and $\mathbf{V}(t)$, it is more convenient to use the distributional representaton 7 for $\mathbf{C}(t)$. Thus, the following steps were followed in the simulation of each observation:

(i) simulate $\mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{I}_{17})$ and $w \sim \chi^2_{20}$, where $\mathbf{Z}$ and W are independent;

(ii) simulate $\mathbf{Z}^* \sim MVN(\mathbf{0}, \mathbf{I}_5)$ and $w^* \sim \chi^2_{20}$, where $\mathbf{Z}^*$ and W^* are independent (and also independent of $\mathbf{Z}$ and W);

TABLE 5

Linear Bayes estimates (Eqn. 16) of source contributions ($\mathbf{H} = 10\mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	4.52	17.08	2.46	0.51	1.24
2	5.00	17.98	3.74	0.43	1.80
3	4.55	24.65	4.12	0.51	1.28
4	5.01	18.69	5.14	0.61	1.29
5	4.07	23.85	4.42	1.15	1.65
6	4.37	18.21	3.71	0.45	1.33
7	5.75	19.38	3.88	0.45	0.98
8	4.58	19.82	2.86	0.35	1.26
9	5.49	20.72	3.35	0.54	1.41
10	4.75	22.06	4.07	0.36	1.70
11	5.35	20.32	4.95	0.80	1.88
12	4.17	17.25	3.27	0.84	1.58
13	4.48	17.13	4.33	0.60	1.04
14	6.01	16.36	2.85	1.19	1.43
15	4.08	28.10	3.77	0.48	1.73
16	5.17	20.21	5.02	0.52	1.52
17	4.00	17.24	3.93	0.34	2.01
18	4.54	17.88	3.26	0.29	1.35
19	4.28	18.04	4.23	0.50	1.54
20	5.58	18.66	4.00	0.52	1.82
21	4.96	23.68	3.40	0.38	1.14
22	5.62	19.19	4.29	0.72	1.80
23	3.85	27.87	3.55	0.36	1.04
24	3.74	22.79	3.89	0.41	1.01
25	4.56	19.68	3.70	0.59	1.75
26	5.52	17.51	3.91	0.43	0.97
27	5.05	16.98	2.98	0.73	1.16
28	4.75	18.44	3.63	0.47	1.49
29	5.07	18.13	3.25	0.46	1.09
30	6.83	18.77	3.53	1.01	1.67
31	4.31	20.55	3.46	0.62	1.39
32	3.83	15.60	2.79	0.82	2.68
33	4.45	18.08	2.92	0.56	1.35
34	4.48	19.11	4.00	0.77	1.66
35	5.12	17.41	4.04	0.76	1.41
36	4.50	25.20	3.38	0.97	1.39
37	5.41	14.99	4.28	0.54	1.58
38	5.35	17.40	5.73	0.53	1.90
39	5.70	19.41	4.69	0.51	1.44
40	5.12	17.08	4.22	0.71	1.61

TABLE 6

Point estimates (Eqn. 12) of source contributions ($\mathbf{H} = \mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.20	2.78	0.08	0.88	1.27
2	7.97	10.03	6.15	0.03	3.91
3	2.95	47.97	10.41	1.40	1.47
4	8.04	15.81	13.93	2.17	1.25
5	5.22	55.28	12.01	4.47	5.12
6	3.91	9.71	6.24	1.02	1.75
7	11.41	25.80	5.48	0.08	0.16
8	4.99	29.61	0.37	0.71	1.42
9	8.80	40.91	3.04	0.50	1.86
10	6.72	35.67	7.51	0.07	3.68
11	8.44	24.76	15.10	2.53	3.90
12	1.74	3.90	5.03	3.26	2.73
13	3.61	29.25	10.90	2.15	0.13
14	12.00	5.15	1.81	5.19	1.55
15	2.88	47.10	10.43	0.71	3.13
16	5.61	25.60	13.03	0.81	2.41
17	0.87	10.56	6.91	0.22	4.85
18	5.68	20.79	0.84	0.13	1.89
19	3.47	9.93	10.21	1.23	2.62
20	10.51	45.78	4.19	0.36	4.47
21	8.49	57.72	5.30	0.51	0.64
22	5.96	30.37	6.07	2.55	3.64
23	0.98	47.26	5.05	0.52	0.95
24	2.39	35.28	7.76	0.75	0.92
25	7.57	52.82	5.46	1.21	4.32
26	12.23	35.98	6.73	0.78	0.27
27	7.43	2.32	3.15	2.25	0.80
28	6.18	13.27	5.66	0.88	2.39
29	8.14	16.61	3.06	0.82	0.58
30	15.22	43.63	2.93	3.52	2.33
31	3.14	39.67	6.10	1.80	1.84
32	4.87	86.74	2.41	1.57	6.55
33	4.58	17.50	2.50	1.31	1.61
34	6.94	47.55	8.27	2.28	3.35
35	8.01	8.52	8.65	2.72	1.74
36	2.00	47.76	7.44	4.48	1.94
37	8.54	23.96	6.84	0.86	2.27
38	9.37	16.59	15.96	1.33	3.95
39	11.18	21.31	10.12	0.51	2.23
40	7.88	6.82	9.53	1.92	2.95

(iii) let $\mathbf{s}(t) = \mu + \sqrt{18}(W^*)^{-1/2}\mathbf{D}^{1/2}\mathbf{Z}^*$; thus $\mathbf{s}(t)$ has a multivariate t-distribution with mean vector μ, degrees of freedom = 20, and covariance matrix $\mathbf{D}$;

(iv) let $\mathbf{C}(t) = \Lambda\mathbf{s}(t) + \left\{\frac{(\nu - m - 2)(1 + \mathbf{s}'(t)\mathbf{H}\mathbf{s}(t))}{[1 + tr(\mathbf{HD}) + \mu'\mathbf{H}\mu]W}\right\}^{1/2} \Psi^{1/2}\mathbf{Z}$.

Since our estimation methods require that the matrix H be specified beforehand, we simulated three data sets of 40 samples of concentration vectors $\mathbf{C}(t)$, $t = 1, \cdots, 40$ for $\mathbf{H} = 10\mathbf{I}_5$, $\mathbf{I}_5$ and $0.1\mathbf{I}_5$, respectively. Note that the values of $\mathbf{s}(t)$ do not depend on the choice of $\mathbf{H}$; the $\mathbf{s}(t)$ were only simulated once for the three data sets of $\mathbf{C}(t)$. The values of $\mathbf{s}(t)$, $t = 1, \cdots, 40$, so produced are given in Table 2.

TABLE 7

Simultaneous 95% confidence intervals (Eqn. 13) for the elements of the source contribution vectors ($\mathbf{H} = \mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.20 ± 0.60	2.78 ± 1.61	0.08 ± 0.48	0.88 ± 0.25	1.27 ± 0.36
2	7.97 ± 1.88	10.03 ± 5.05	6.15 ± 1.50	0.03 ± 0.78	3.91 ± 1.12
3	2.95 ± 5.26	47.97 ± 14.15	10.41 ± 4.20	1.40 ± 2.19	1.47 ± 3.15
4	8.04 ± 2.49	15.81 ± 6.69	13.93 ± 1.99	2.17 ± 1.04	1.25 ± 1.49
5	5.22 ± 5.93	55.28 ± 15.95	12.01 ± 4.74	4.47 ± 2.47	5.12 ± 3.55
6	3.91 ± 1.26	9.71 ± 3.39	6.24 ± 1.01	1.02 ± 0.52	1.75 ± 0.75
7	11.41 ± 3.11	25.80 ± 8.37	5.48 ± 2.49	0.08 ± 1.30	0.16 ± 1.86
8	4.99 ± 3.70	29.61 ± 9.96	0.37 ± 2.96	0.71 ± 1.54	1.42 ± 2.22
9	8.80 ± 4.94	40.91 ± 13.28	3.04 ± 3.95	0.50 ± 2.06	1.86 ± 2.95
10	6.72 ± 4.80	35.67 ± 12.90	7.51 ± 3.83	0.07 ± 2.00	3.68 ± 2.87
11	8.44 ± 3.87	24.76 ± 10.41	15.10 ± 3.09	2.53 ± 1.61	3.90 ± 2.32
12	1.74 ± 1.01	3.90 ± 2.72	5.03 ± 0.81	3.26 ± 0.42	2.73 ± 0.60
13	3.61 ± 3.47	29.25 ± 9.34	10.90 ± 2.78	2.15 ± 1.45	0.13 ± 2.08
14	12.00 ± 1.63	5.15 ± 4.38	1.81 ± 1.30	5.19 ± 0.68	1.55 ± 0.97
15	2.88 ± 5.59	47.10 ± 15.04	10.43 ± 4.47	0.71 ± 2.33	3.13 ± 3.35
16	5.61 ± 3.24	25.60 ± 8.71	13.03 ± 2.59	0.81 ± 1.35	2.41 ± 1.94
17	0.87 ± 1.40	10.56 ± 3.77	6.91 ± 1.12	0.22 ± 0.58	4.85 ± 0.84
18	5.68 ± 1.20	20.79 ± 3.22	0.84 ± 0.96	0.13 ± 0.50	1.89 ± 0.72
19	3.47 ± 1.83	9.93 ± 4.91	10.21 ± 1.46	1.23 ± 0.76	2.62 ± 1.09
20	10.51 ± 5.71	45.78 ± 15.36	4.19 ± 4.57	0.36 ± 2.38	4.47 ± 3.42
21	8.49 ± 7.47	57.72 ± 20.09	5.30 ± 5.97	0.51 ± 3.11	0.64 ± 4.47
22	5.96 ± 1.64	30.37 ± 4.41	6.07 ± 1.31	2.55 ± 0.68	3.64 ± 0.98
23	0.98 ± 5.73	47.26 ± 15.41	5.05 ± 4.58	0.52 ± 2.39	0.95 ± 3.43
24	2.39 ± 3.66	35.28 ± 9.84	7.76 ± 2.93	0.75 ± 1.52	0.92 ± 2.19
25	7.57 ± 6.20	52.82 ± 16.69	5.46 ± 4.96	1.21 ± 2.59	4.32 ± 3.71
26	12.23 ± 4.62	35.98 ± 12.44	6.73 ± 3.70	0.78 ± 1.93	0.27 ± 2.77
27	7.43 ± 0.95	2.32 ± 2.57	3.15 ± 0.76	2.25 ± 0.40	0.80 ± 0.57
28	6.18 ± 1.91	13.27 ± 5.15	5.66 ± 1.53	0.88 ± 0.80	2.39 ± 1.14
29	8.14 ± 2.20	16.61 ± 5.93	3.06 ± 1.76	0.82 ± 0.92	0.58 ± 1.32
30	15.22 ± 5.35	43.63 ± 14.38	2.93 ± 4.27	3.52 ± 2.23	2.33 ± 3.20
31	3.14 ± 5.17	39.67 ± 13.89	6.10 ± 4.13	1.80 ± 2.15	1.84 ± 3.09
32	4.87 ± 6.75	86.74 ± 18.16	2.41 ± 5.40	1.57 ± 2.81	6.55 ± 4.04
33	4.58 ± 2.24	17.50 ± 6.03	2.50 ± 1.79	1.31 ± 0.93	1.61 ± 1.34
34	6.94 ± 5.65	47.55 ± 15.19	8.27 ± 4.51	2.28 ± 2.35	3.35 ± 3.38
35	8.01 ± 1.62	8.52 ± 4.36	8.65 ± 1.29	2.72 ± 0.67	1.74 ± 0.97
36	2.00 ± 5.71	47.76 ± 15.37	7.44 ± 4.57	4.48 ± 2.38	1.94 ± 3.42
37	8.54 ± 2.60	23.96 ± 6.99	6.84 ± 2.08	0.86 ± 1.08	2.27 ± 1.56
38	9.37 ± 3.19	16.59 ± 8.57	15.96 ± 2.55	1.33 ± 1.33	3.95 ± 1.91
39	11.18 ± 3.24	21.31 ± 8.71	10.12 ± 2.59	0.51 ± 1.35	2.23 ± 1.94
40	7.88 ± 1.85	6.82 ± 4.97	9.53 ± 1.48	1.92 ± 0.77	2.95 ± 1.11

We showed in the beginning of this section how the measured concentration vectors $\mathbf{C}(1), \cdots, \mathbf{C}(40)$ can be used to estimate the values of the parameters $\boldsymbol{\Lambda}$, μ $\mathbf{D}$, ν, $\mathbf{B}$ and $\boldsymbol{\Psi}$ needed in computing the estimates 12 and 16 of the source contributions. As a simplification, rather than actually implementing TTFA to get these estimators, we have assumed that these parameters are known to have the values used to simulate the source contribution vectors $\mathbf{s}(1), \cdots, \mathbf{s}(40)$ and the concentration vectors $\mathbf{C}(1), \cdots, \mathbf{C}(40)$.

Tables 3, 6 and 9 give the estimates 12 of the simulated source contributions, $\mathbf{s}(t)$, $t = 1, \cdots, 40$, with $\mathbf{H}$ being chosen as $10\mathbf{I}_5$, $\mathbf{I}_5$ and $0.1\mathbf{I}_5$, respec-

TABLE 8

Linear Bayes estimates (Eqn. 16) of source contributions ($\mathbf{H} = \mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	4.52	17.08	2.46	0.51	1.24
2	5.00	17.99	3.74	0.43	1.80
3	4.55	24.65	4.12	0.51	1.28
4	5.01	18.69	5.14	0.61	1.29
5	4.07	23.84	4.42	1.15	1.65
6	4.37	18.22	3.71	0.45	1.33
7	5.75	19.38	3.88	0.45	0.98
8	4.58	19.82	2.86	0.35	1.26
9	5.49	20.72	3.35	0.54	1.41
10	4.75	22.06	4.07	0.36	1.70
11	5.35	20.32	4.95	0.80	1.88
12	4.17	17.25	3.27	0.84	1.58
13	4.48	17.13	4.33	0.60	1.04
14	6.01	16.36	2.85	1.19	1.43
15	4.08	28.09	3.77	0.48	1.73
16	5.17	20.21	5.02	0.52	1.52
17	4.00	17.24	3.93	0.34	2.01
18	4.54	17.88	3.26	0.29	1.35
19	4.28	18.04	4.22	0.50	1.54
20	5.58	18.66	4.00	0.52	1.82
21	4.96	23.68	3.40	0.38	1.14
22	5.62	19.19	4.29	0.72	1.80
23	3.85	27.87	3.55	0.36	1.04
24	3.74	22.79	3.89	0.41	1.01
25	4.56	19.68	3.70	0.59	1.75
26	5.52	17.51	3.91	0.43	0.97
27	5.05	16.98	2.98	0.73	1.16
28	4.75	18.44	3.63	0.47	1.49
29	5.07	18.13	3.25	0.46	1.09
30	6.83	18.78	3.53	1.01	1.67
31	4.31	20.55	3.46	0.62	1.39
32	3.83	15.61	2.79	0.82	2.68
33	4.45	18.08	2.92	0.56	1.35
34	4.48	19.11	4.00	0.77	1.66
35	5.12	17.41	4.04	0.76	1.41
36	4.50	25.20	3.38	0.97	1.39
37	5.41	14.99	4.28	0.54	1.58
38	5.35	17.40	5.73	0.53	1.90
39	5.70	19.41	4.69	0.51	1.44
40	5.12	17.08	4.22	0.71	1.61

TABLE 9

Point estimates (Eqn. 12) of source contributions ($\mathbf{H} = 0.1\mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.19	2.73	0.08	0.86	1.27
2	7.96	10.00	6.15	0.03	3.91
3	2.95	47.88	10.41	1.41	1.47
4	8.04	15.77	13.93	2.18	1.25
5	5.21	55.16	12.01	4.47	5.12
6	3.90	9.73	6.22	1.02	1.76
7	11.41	25.74	5.48	0.08	0.16
8	4.99	29.54	0.38	0.71	1.42
9	8.80	40.83	3.04	0.50	1.86
10	6.71	35.60	7.51	0.07	3.68
11	8.44	24.71	15.09	2.53	3.90
12	1.75	3.86	5.04	3.27	2.70
13	3.61	29.19	10.89	2.15	0.14
14	12.00	5.11	1.81	5.19	1.55
15	2.88	46.98	10.43	0.72	3.13
16	5.61	25.55	13.02	0.81	2.41
17	0.88	10.53	6.91	0.23	4.85
18	5.68	20.74	0.84	0.14	1.89
19	3.46	9.90	10.21	1.23	2.63
20	10.50	45.67	4.20	0.37	4.46
21	8.49	57.59	5.31	0.51	0.64
22	5.95	30.31	6.07	2.55	3.64
23	0.98	47.15	5.05	0.52	0.95
24	2.39	35.21	7.76	0.75	0.92
25	7.57	52.70	5.46	1.21	4.31
26	12.22	35.90	6.73	0.79	0.27
27	7.43	2.36	3.15	2.25	0.82
28	6.17	13.22	5.66	0.89	2.40
29	8.14	16.56	3.06	0.82	0.58
30	15.21	43.51	2.94	3.52	2.33
31	3.14	39.57	6.10	1.80	1.84
32	4.86	86.56	2.41	1.58	6.55
33	4.58	17.46	2.50	1.31	1.61
34	6.95	47.43	8.27	2.29	3.35
35	8.00	8.50	8.66	2.71	1.74
36	1.99	47.66	7.45	4.47	1.94
37	8.54	23.91	6.84	0.86	2.27
38	9.37	16.56	15.96	1.33	3.95
39	11.17	21.26	10.12	0.52	2.24
40	7.89	6.80	9.53	1.92	2.95

tively. The simultaneous confidence intervals 13 for the elements of $\mathbf{s}(t)$, $t = 1, \cdots, 40$, are presented in tables 4, 7 and 10. Finally, the linear Bayes estimates 16 of the source contributions are reported in Tables 5, 8 and 11 for each of the three different choices of $\mathbf{H}$.

As an illustration, we give the point estimates 12 and 16 and the 95% confidence intervals 13 for source contribution vector $\mathbf{s}(14)$ from the 14th sample when $\mathbf{H} = 10\mathbf{I}_5$. The true source contribution vector for the 14th sample is given in the 14th row of table 2 as

$$\mathbf{s}(14) = (11.13, 5.75, 1.85, 5.11, 1.59)'$$

The estimates of $\mathbf{s}(14)$ given by Eqn. 12 and by

TABLE 10

Simultaneous 95% confidence intervals (Eqn. 13) for the elements of the source contribution vectors ($\mathbf{H} = 0.1\mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	5.19 ± 0.66	2.73 ± 1.78	0.08 ± 0.53	0.86 ± 0.28	1.27 ± 0.40
2	7.96 ± 1.90	10.00 ± 5.11	6.15 ± 1.52	0.03 ± 0.79	3.91 ± 1.14
3	2.95 ± 5.22	47.88 ± 14.04	10.41 ± 4.18	1.41 ± 2.18	1.47 ± 3.13
4	8.04 ± 2.49	15.77 ± 6.68	13.93 ± 1.99	2.18 ± 1.04	1.25 ± 1.49
5	5.21 ± 5.88	55.16 ± 15.81	12.01 ± 4.70	4.47 ± 2.45	5.12 ± 3.52
6	3.90 ± 1.28	9.73 ± 3.45	6.22 ± 1.03	1.02 ± 0.54	1.76 ± 0.77
7	11.41 ± 3.10	25.74 ± 8.34	5.48 ± 2.48	0.08 ± 1.29	0.16 ± 1.86
8	4.99 ± 3.69	29.54 ± 9.92	0.38 ± 2.95	0.71 ± 1.54	1.42 ± 2.21
9	8.80 ± 4.91	40.83 ± 13.18	3.04 ± 3.92	0.50 ± 2.04	1.86 ± 2.94
10	6.71 ± 4.77	35.60 ± 12.82	7.51 ± 3.81	0.07 ± 1.99	3.68 ± 2.85
11	8.44 ± 3.86	24.71 ± 10.36	15.09 ± 3.08	2.53 ± 1.61	3.90 ± 2.31
12	1.75 ± 1.07	3.86 ± 2.88	5.04 ± 0.86	3.27 ± 0.45	2.70 ± 0.64
13	3.61 ± 3.46	29.19 ± 9.29	10.89 ± 2.76	2.15 ± 1.44	0.14 ± 2.07
14	12.00 ± 1.65	5.11 ± 4.44	1.81 ± 1.32	5.19 ± 0.69	1.55 ± 0.99
15	2.88 ± 5.55	46.98 ± 14.91	10.43 ± 4.44	0.72 ± 2.31	3.13 ± 3.32
16	5.61 ± 3.23	25.55 ± 8.67	13.02 ± 2.58	0.81 ± 1.34	2.41 ± 1.93
17	0.88 ± 1.42	10.53 ± 3.82	6.91 ± 1.14	0.23 ± 0.59	4.85 ± 0.85
18	5.68 ± 1.20	20.74 ± 3.21	0.84 ± 0.96	0.14 ± 0.50	1.89 ± 0.72
19	3.46 ± 1.85	9.90 ± 4.97	10.21 ± 1.48	1.23 ± 0.77	2.63 ± 1.11
20	10.50 ± 5.67	45.67 ± 15.24	4.20 ± 4.54	0.37 ± 2.36	4.46 ± 3.39
21	8.49 ± 7.41	57.59 ± 19.92	5.31 ± 5.93	0.51 ± 3.09	0.64 ± 4.44
22	5.95 ± 1.63	30.31 ± 4.38	6.07 ± 1.30	2.55 ± 0.68	3.64 ± 0.98
23	0.98 ± 5.69	47.15 ± 15.29	5.05 ± 4.55	0.52 ± 2.37	0.95 ± 3.40
24	2.39 ± 3.64	35.21 ± 9.78	7.76 ± 2.91	0.75 ± 1.52	0.92 ± 2.18
25	7.57 ± 6.16	52.70 ± 16.54	5.46 ± 4.92	1.21 ± 2.57	4.31 ± 3.68
26	12.22 ± 4.60	35.90 ± 12.36	6.73 ± 3.68	0.79 ± 1.92	0.27 ± 2.75
27	7.43 ± 1.00	2.36 ± 2.69	3.15 ± 0.80	2.25 ± 0.42	0.82 ± 0.60
28	6.17 ± 1.93	13.22 ± 5.19	5.66 ± 1.55	0.89 ± 0.81	2.40 ± 1.16
29	8.14 ± 2.21	16.56 ± 5.95	3.06 ± 1.77	0.82 ± 0.92	0.58 ± 1.33
30	15.21 ± 5.31	43.51 ± 14.26	2.94 ± 4.24	3.52 ± 2.21	2.33 ± 3.17
31	3.14 ± 5.13	39.57 ± 13.79	6.10 ± 4.10	1.80 ± 2.14	1.84 ± 3.07
32	4.86 ± 6.68	86.56 ± 17.96	2.41 ± 5.34	1.58 ± 2.78	6.55 ± 4.00
33	4.58 ± 2.25	17.46 ± 6.05	2.50 ± 1.80	1.31 ± 0.94	1.61 ± 1.35
34	6.95 ± 5.60	47.43 ± 15.06	8.27 ± 4.48	2.29 ± 2.34	3.35 ± 3.35
35	8.00 ± 1.64	8.50 ± 4.40	8.66 ± 1.31	2.71 ± 0.68	1.74 ± 0.98
36	1.99 ± 5.67	47.66 ± 15.25	7.45 ± 4.54	4.47 ± 2.36	1.94 ± 3.40
37	8.54 ± 2.59	23.91 ± 6.96	6.84 ± 2.07	0.86 ± 1.08	2.27 ± 1.55
38	9.37 ± 3.18	16.56 ± 8.55	15.96 ± 2.55	1.33 ± 1.33	3.95 ± 1.90
39	11.17 ± 3.23	21.26 ± 8.69	10.12 ± 2.58	0.52 ± 1.35	2.24 ± 1.93
40	7.89 ± 1.87	6.80 ± 5.03	9.53 ± 1.50	1.92 ± 0.78	2.95 ± 1.12

the linear Bayes estimator 16 are obtained from Tables 3 and 5, respectively, as

$$\hat{\mathbf{s}}(14) = (12.00, 5.15, 1.81, 5.19, 1.55)'$$

$$\tilde{\mathbf{s}}(14) = (6.01, 16.38, 2.85, 1.19, 1.43)'$$

TABLE 11

Linear Bayes estimates (Eqn. 16) of source contributions ($\mathbf{H} = 0.1\mathbf{I}_5$)

Samples	Motor vehicle	Sulfate	Flyash/ soil	Paint	Refuse
1	4.51	17.10	2.44	0.51	1.24
2	4.99	18.00	3.74	0.43	1.80
3	4.54	24.64	4.12	0.51	1.28
4	5.01	18.69	5.14	0.61	1.29
5	4.08	23.83	4.42	1.15	1.65
6	4.37	18.24	3.71	0.45	1.33
7	5.75	19.39	3.88	0.45	0.98
8	4.58	19.81	2.86	0.35	1.26
9	5.48	20.72	3.35	0.54	1.41
10	4.75	22.06	4.07	0.36	1.70
11	5.35	20.32	4.94	0.80	1.88
12	4.17	17.29	3.27	0.83	1.58
13	4.48	17.13	4.33	0.60	1.04
14	6.01	16.35	2.86	1.19	1.43
15	4.08	28.06	3.77	0.48	1.74
16	5.17	20.21	5.02	0.52	1.52
17	4.01	17.24	3.93	0.34	2.01
18	4.54	17.87	3.26	0.29	1.35
19	4.27	18.05	4.22	0.50	1.54
20	5.57	18.67	4.00	0.52	1.82
21	4.96	23.66	3.40	0.38	1.14
22	5.62	19.19	4.29	0.72	1.80
23	3.85	27.84	3.55	0.36	1.04
24	3.74	22.78	3.89	0.42	1.01
25	4.56	19.70	3.71	0.59	1.75
26	5.52	17.51	3.91	0.43	0.97
27	5.05	17.00	2.97	0.73	1.16
28	4.75	18.44	3.63	0.47	1.49
29	5.07	18.12	3.25	0.46	1.09
30	6.83	18.80	3.53	1.01	1.66
31	4.31	20.54	3.46	0.62	1.39
32	3.82	15.67	2.78	0.82	2.68
33	4.45	18.07	2.92	0.56	1.35
34	4.48	19.11	4.00	0.77	1.66
35	5.12	17.41	4.04	0.76	1.41
36	4.50	25.17	3.38	0.96	1.38
37	5.41	14.99	4.28	0.54	1.58
38	5.35	17.40	5.73	0.53	1.90
39	5.70	19.41	4.69	0.51	1.44
40	5.12	17.09	4.22	0.71	1.61

From Table 4, the simultaneous 95% confidence intervals for the elements of $\mathbf{s}(14)$ are

$$\begin{aligned}
&(10.37,\ 13.63),\\
&(0.77,\ 9.53),\\
&(0.51,\ 3.11),\\
&(4.51,\ 5.87),\\
&(0.58,\ 2.52).
\end{aligned}$$

Note, for this sample, that the estimates 12 give very precise approximations to the source contributions and that the elements of $\mathbf{s}(14)$ are contained in the simultaneous confidence intervals 13.

DISCUSSION

The model of the present paper constitutes our initial attempt to account for the criticisms, given in the introductory part, of current models and methods of aerosol mass apportionment. It provides two point estimators, Eqns. 12 and 16, for the vectors of source contributions $\mathbf{s}(t)$. A simulation study reveals that both methods give reasonable approximations to the true source contribution vectors. Most important, this paper also offers a way to evaluate the accuracy of estimation of $\mathbf{s}(t)$ through the use of simultaneous confidence intervals.

REFERENCES

1 R.C. Henry, C.W. Lewis, P.K. Hopke and H.J. Williamson, Atmos. Environ., 18 (1984) 1507.
2 P.K. Hopke, Receptor Modeling in Environmental Chemistry, Wiley, New York, 1985.
3 J.G. Watson, Chemical Element Balance Receptor Methodology for Assessing the Sources of Fine and Total Suspended Particulate Matter in Portland, Oregon, Ph.D. Dissertation, Oregon Graduate Center, Beaverton, OR, 1979.
4 J.G. Watson, J.A. Cooper and J.H. Huntzicker, Atmos. Environ., 18 (1984) 1347.
5 M.D. Cheng, P.K. Hopke and D.E. Jennings, Chemom. Intell. Lab. Syst., 4 (1988) 239.
6 P.K. Hopke, Atmos. Environ., 22 (1988) 1777.
7 D.L. Duewer, B.R. Kowalski and J.L. Fasching, Anal. Chem., 48 (1976) 2002.

8 P.K. Hopke, R.E. Lamb and D.F. Natusch, Environ. Sci. Technol., 14 (1980) 164
9 P.K. Hopke, E.S. Gladney, G.E. Gorden, W.H. Zoller and A.G. Jones, Atmos. Environ., 10 (1976) 1015.
10 H.F. Kaiser, Education and Psychological Measurement, 19 (1959) 413.
11 T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 2nd edn., 1984.
12 L.J. Gleser and J.T. Hwang, Ann. Stat., 15 (1987) 1351.
13 M.D. Cheng and P.K. Hopke, Chemom. Intell. Lab. Syst., 1 (1986) 33.
14 L.A. Currie, Chemom. Intell. Lab. Syst., 10 (1991) 59.
15 D.J. Albert and P.K. Hopke, Atmos. Environ., 15 (1981) 675.
16 S.J. Press, Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, Krieger Publ., Malabar, FL, 1982.
17 C. Morris, J. Am. Stat. Assoc., 78 (1983) 47.
18 K.G. Joreskög and D. Sörbum, LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variable, and Least Squares Methods, Scientific Software, Inc., Mooresvile, IN, 1986.
19 B.P. Carlin and A.E. Gelfand, J. Am. Stat. Assoc., 85 (1990) 105.

Analytica Chimica Acta, 277 (1993) 421–429
Elsevier Science Publishers B.V., Amsterdam

Receptor models to study groundwater contamination

Edward J. Baum

Department of Chemistry, Grand Valley State University, Allendale, MI 49401 (USA)

(Received 3rd September 1992)

Abstract

This paper describes the use of receptor modeling to accomplish three tasks: estimating the number of sources making significant contributions to groundwater contamination at a receptor site, revealing the chemical signatures of effluent from each pollutant source, and apportioning responsibility for contamination at a receptor site among the sources. Physical and chemical processes cannot be ignored in groundwater modeling as they usually are in air pollution receptor modeling. If the contaminants undergo differential chemical reactivity and retardation in the soil column, estimates of the number of pollution sources will be inflated by the number of important physical chemical processes. This problem is addressed by coupling pattern recognition and factor analysis methods. Chemical reactivity and retardation cause changes in apparent source signatures with time and contribute to error in source apportionment with linear receptor models. The receptor modeling process reveals the nature and rates of physical chemical processes occurring in the soil column, allowing correction of apparent source signatures. Model error can be reduced by use of non-linear receptor models employing corrected source signatures.

Keywords: Groundwater; Receptor models; Waters

Receptor models are empirical relationships that predict the emission characteristics of contaminant sources based on the spatial and temporal distribution of contaminant levels observed at receptor sites. They are widely used in the study of air pollution, but they have not yet been applied to the study of groundwater problems. This paper explores the use of receptor modeling to study groundwater pollution.

As several recent reviews point out [1–3], receptor models are widely used to determine the impact of suspended particulate sources on air quality. One frequently cited example of the effective use of receptor modeling is the aerosol study conducted in Portland, OR, in which receptor modeling was used to identify and estimate the impact on air quality of sources of atmospheric aerosol and lead. An unexpected source, reentrained street dust, was shown to be a major contributor of both suspended particulate and particulate lead [4]. Field measurements of the chemical signatures of the identified sources led to several results: improved apportionment of source contributions [5,6] chemical characterization of air pollution regimes in the airshed [7], and the correction and calibration of a simulation model used to make air quality management decisions [6]. The study resulted in major changes in Portland's air quality control strategy, dramatically improving its effectiveness.

Informed and rational groundwater quality management decisions are based, in part, on simulation model estimates of contaminant transport in the soil column. However, simulation models have several shortcomings. Simulation models are quantitative mathematical relationships that employ source emission inventories and dispersion models to describe the transport and fate of

Correspondence to: E.J. Baum, Department of Chemistry, Grand Valley State University, Allendale, MI 49401 (USA).

contaminants in the soil column. All important sources, transport mechanisms and transformations of contaminants must be accurately represented if the model is to produce accurate and useful estimates.

But complete information about groundwater systems is rarely available. Uncertainties in groundwater flow patterns and in the rates of emission, dispersion, and chemical conversion reduce the ability of simulation models to predict the effect of control strategies. No current model can estimate non-equilibrium transport of contaminants and transport through structured soils containing voids and macropores, conditions that prevail in real soil columns. Few models are available that satisfactorily estimate solute transport in multiple liquid phases.

Furthermore, subsurface flow patterns can be complex, putting the relative location of the source in doubt. It is possible to misidentify sources when there is a number of choices. Some sources may be transient, hidden, and unsuspected. In the absence of painstaking field work, it is likely that the location and strength of some important contaminant sources will be unknown. In such cases, remediation will not achieve the anticipated changes in water quality at a receptor site.

These problems can be avoided through the use of receptor modeling, as this report demonstrates. The report describes receptor modeling as applied to study a typical problem involving contamination of a municipal water supply. Data on the spatial and temporal variation of groundwater contamination are used to estimate the number of sources making measurable contributions to contamination at the receptor sites, reveal the chemical signatures of effluent from each source, and apportion responsibility for contamination at a receptor site among the sources.

RECEPTOR SITE MONITORING DATA

The municipal water supply of the City of Battle Creek, MI, is supplied by a well field located in a heavily industrialized area northeast of the city. The well field's principal aquifer is a sandstone formation that lies about 25 m below the surface. The soil column above the formation is porous glacial till.

The well field was found to be contaminated with industrial solvents and degreasers in August, 1981. Well-water levels of tetrachloroethylene, trichloroethylene, and 1,2-dichloroethylene in the parts-per-million (ppm) range were found in the southern part of the well field. Subsequent work by the Calhoun County Health Department, the Michigan Department of Public Health, the Michigan Department of Natural Resources, the U.S. Geological Survey, and the U.S. Environmental Protection Agency established a grid of monitoring and purge wells to the south of the well field, determined the general groundwater flow patterns (from southeast to northwest) and flow-rates in the area (275 meters per year), and identified two potential sources of contamination [8]. The suspected sources lie 300 m apart along an east-west line located 600 m south of the well field.

Several factors caused uncertainty in the results of the conventional hydrogeological studies performed at the site. Over time, land use and potential contaminant sources in the area changed. Also, groundwater flow patterns changed dramatically during the study period as a result of pollution control efforts. No historic data are available to account for these factors.

Finally, two organic liquid phases may exist in the soil column in addition to the aqueous phase. One is a predominantly aromatic hydrocarbon phase lighter than water, and one is a predominantly halocarbon phase heavier than water. An organic liquid phase was found floating on the groundwater at site A, and other liquid phases may exist in the soil column, as well. There is evidence that a dense halocarbon plume would travel by gravity flow down the slope of bedrock in a direction different from that of the groundwater.

The data used in the receptor model study consists of contaminant levels in 177 well-water samples taken by the NUS Co. and analyzed for 20 organic pollutants using standard U.S. EPA procedures. The data set was described in a previous report [8]. The data were collected between

TABLE 1

Average well-water contaminant concentration

Analyte	Mean (ppb)	S.D. (ppb)
1 Acetone	9107.1	76693.8
2 2-Butanone	656.4	4131.3
3 4-Methyl-2-pentanone	293.0	2112.2
4 Benzene	13668.4	128594.0
5 Ethylbenzene	52293.3	528602.3
6 Toluene	511791.8	4757196.0
7 Xylene	164617.0	1835317.0
8 Methylene chloride	19467.0	207020.0
9 Chloroform	364.2	4497.6
10 Fluorotrichloromethane	1.3	8.1
11 1,1,1-Trichloroethane	133485.0	1516589.0
12 1,2-Dichloroethane	7739.6	97428.3
13 1,1-Dichloroethane	296.7	3157.4
14 Tetrachloroethylene	237324.0	2665221.0
15 Trichloroethylene	269067.0	2768089.0
16 1,2-Dichloroethylene	2466.6	21265.0
17 1,1-Dichloroethylene	7982.2	104928.0
18 Vinyl chloride	21.2	182.6
19 Volatile organics	1432212.0	14017215.0

January, 1983, and December, 1987, in order to monitor groundwater contamination, and it is typical of such efforts. A large number of records are censored, and only a portion is useable. Two compounds, 1,2-dichloropropane and trichlorotrifluoroethane, were found to be below detection limits in all samples and were not included in the data analysis. Three, acetone, 2-butanone and 4-methyl-2-pentanone, appeared in some samples due to contamination during handling in the laboratory, and these were not included in the data analysis. Methylene chloride, chloroform and vinyl chloride were found above detection limits in only a few samples. Omission of these did not alter the results of the data analysis. Summary statistics are given in Table 1.

Even though the data set contains a high percentage of censored values, the parameter frequency distributions appear lognormal. In all cases, a log transform was applied in order to produce normally distributed data sets for statistical analysis. No outliers were identified using pattern recognition methods.

RECEPTOR MODELING

The fundamental mathematical relationship of receptor modeling describes the conservation of mass of contaminant as it is transported from source(s) to receptor site(s). The relationship is:

$$\mathbf{C} = \mathbf{SM} + \mathbf{E} \quad (1)$$

where **C** is the matrix of contaminant concentration profiles found in samples taken at the receptor site(s), **S** is the matrix of measured or estimated source signatures (profiles of contaminant mass fraction in source emissions), **M** is the source strength matrix to be calculated (profiles of the mass contributions of each source to each sample), and **E** is the measurement error matrix. Depending on the way in which **C**, **M**, and **E** are defined (i.e., whether they are scalars, vectors, or matrices), the calculation procedure may be called chemical mass balance, multiple linear regression, or factor analysis. In any event, the procedure involves:

(1) determining the number of sources making measurable contributions to contaminant levels at the receptor sites;

(2) constructing the matrix of chemical signatures, **S**, characteristic of effluent from each source; and

(3) apportioning responsibility for contamination observed at the receptor site(s) among the sources by solving Eqn. 1 for **M**.

Determining the number of significant sources

The number of important controls of groundwater chemistry is usually assumed to be equal to the inherent dimensionality of the data set. The dimensionality is estimated by analysis of goodness-of-fit statistics calculated in attempting to reproduce the data set with varying numbers of derived factors.

Since there is a wide variation in average concentration among the analytes, a Q-mode factor analysis was performed in order to equally weight all analytes in the analysis. The computer program FANTASIA, originally created for aerosol source apportionment, was used [9]. The statistics used were the eigenvalue of the correlation matrix, the root-mean-square error (RMS), the chi-

square statistic, the Exner function, the Indicator function, and the average error. These and other goodness-of-fit statistics are described and their use discussed by Jolliffe [10] and by Malinowski and Howery [11].

The analysis and data reproduction results, shown in Table 2, suggest that there are between three and seven significant sources contributing to groundwater contamination at the study site. The indicator function reaches a minimum at three factors, whereas trends of the eigenvalues and goodness-of-fit parameters suggest that as many as seven factors best reproduce the data set.

The divergence between indicator, eigenvalue, and goodness-of-fit estimates may be due to changes in source signatures with time, the existence of important analyte removal processes, and differential retardation in the soil column. Important physical and chemical processes generate variance in the data, which results in the extraction of additional linearly independent factors in the factor analysis and inflation of the estimated number of sources.

The issue can be resolved by use of factor score plots, also known as Karhunen-Loeve transformations and Hotelling transformations, to examine the data. This pattern recognition method was used by Wegscheider and Leyden [12] to classify groundwater samples taken in Colorado. Meglen and Sistko [13], in a study of groundwater samples taken in western Colorado, used factor score plots to classify water from different aquifers and to gain some insight into water chemistry. Usinoff and Guzman-Guzman [14] used factor score plots and correspondence analysis along with traditional *R*- and *Q*-mode factor analysis to study the groundwater chemistry of the Milk River aquifer in Alberta, Canada, and of the San Pedro River basin in Arizona.

Figure 1 shows a three-dimensional projection of the data set obtained by plotting the factor scores of each sample for the first three factors obtained in the *Q*-mode analysis. About 76% of the variance in the twelve-dimensional data set is accounted for by the first three factors. Although there is some overlap, the data appear to be divided primarily into three clusters, labeled A, B, and C, symmetrically arranged about the origin.

The Euclidian distance between data points in Fig. 1 is a quantitative, inverse measure of the chemical similarity between samples. Clustering in the data projection shows that there are three groups of wells with common contaminant concentration profiles. Two of the clusters, A and C, include samples taken from the two sources uncovered in previous hydrogeological studies along with samples from wells that are presumed to be affected by their contaminant plumes. The third cluster, B, includes samples from a group of wells geographically located between those represented

TABLE 2

Q-Mode factor analysis results and reproduction summary for the groundwater data set

Factor No.	Eigenvalue	RMS	Chi-square [a]	Exner	Indicator	Average error (%)
1	31.347	252.67	$5.3501 \cdot 10^{7}$	0.830035	$1.7663 \cdot 10^{-3}$	269.51
2	13.197	214.14	$4.2964 \cdot 10^{7}$	0.633803	$1.7116 \cdot 10^{-3}$	162.26
3	6.9152	95.079	$9.5679 \cdot 10^{6}$	0.501192	$1.7613 \cdot 10^{-3}$	158.23
4	3.9733	61.861	$4.6337 \cdot 10^{6}$	0.405847	$1.9146 \cdot 10^{-3}$	90.37
5	2.4160	54.452	$4.1740 \cdot 10^{6}$	0.334854	$2.2057 \cdot 10^{-3}$	88.86
6	2.3505	47.009	$3.6931 \cdot 10^{6}$	0.246015	$2.3912 \cdot 10^{-3}$	67.26
7	1.3529	43.816	$3.9188 \cdot 10^{6}$	0.177540	$2.7122 \cdot 10^{-3}$	60.76
8	0.75622	30.115	$2.3561 \cdot 10^{6}$	0.122723	$3.2751 \cdot 10^{-3}$	43.15
9	0.50629	16.323	$9.3997 \cdot 10^{5}$	0.063570	$3.4825 \cdot 10^{-3}$	18.44
10	0.096976	13.849	$1.0341 \cdot 10^{6}$	0.043936	$6.6327 \cdot 10^{-3}$	14.24
11	0.076219	4.7468	$2.4765 \cdot 10^{5}$	0.016475	$1.4069 \cdot 10^{-2}$	4.98

[a] Chi-square not weighted by errors.

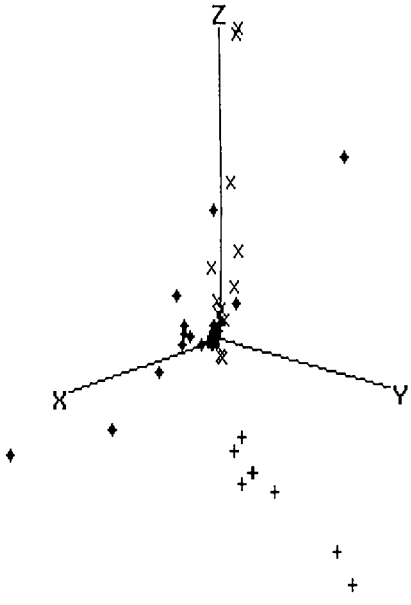


Fig. 1. Q-Mode factor projection of the receptor site data. Sample scores are plotted for the first three factors extracted from the autoscaled data set.

in clusters A and C. The presence of a third major contaminant plume is indicated, although the plume may be due to one of the several liquid phases released by sources A and B.

Feature loadings for the first three factors are shown in Table 3. Factor 1 (X-axis) scores are correlated with groundwater levels of most analytes except ethylbenzene, 1,1-dichloroethane, and the dichloroethylenes. Factor 2 (Y-axis) scores are correlated with concentration levels of benzene, the dichloroethanes, and the dichlorothylenes. Factor 3 (Z-axis) scores are correlated with levels of ethylbenzene and 1,2-dichloroethylene, and they are inversely correlated with levels of 1,1-dichloroethane and 1,1-dichloroethylene.

The feature loadings and elongation of the data clusters are due, in part, to changes in apparent source signatures with time. The extreme data points in clusters A and C describe the earliest samples taken in the vicinity of the presumed sources. The extreme points in all clusters describe samples taken from wells that are furthest upgradient of all those in the cluster.

Chemical conversion alters apparent source signatures as contaminants are transported from source to receptor. The aromatic hydrocarbons undergo microbial-mediated reactions to form oxygenated byproducts with half-lives of years (benzene and ethylbenzene) to months (toluene and xylene) [15,16]. The chlorinated hydrocarbons undergo reductive dehalogenation [17,18] in a sequence leading from tetrachloroethylene, trichloroethylene, and the dichloroethanes to the dichloroethylenes and vinyl chloride [19]. The conversion rates are reported to be small with half-lives, $\tau_{1/2}$, extending from months to years [20]. The half-lives of contaminants measured in this study are an order of magnitude larger than those reported in the literature. Chemical conver-

TABLE 3

Factor loadings of analytes

Analyte	Factors 1 (X-axis)	Factor 2 (Y-axis)	Factor 3 (Z-axis)
4 Benzene	0.269	0.401	0.028
5 Ethylbenzene	0.013	0.071	0.595
6 Toluene	0.413	−0.028	−0.061
7 Xylene	0.387	−0.126	0.046
8 Methylene chloride	0.390	0.016	0.011
11 1,1,1-Trichloroethane	0.402	−0.121	−0.099
12 1,2-Dichloroethane	0.213	0.365	0.081
13 1,1-Dichloroethane	−0.038	0.371	−0.491
14 Tetrachloroethylene	0.283	−0.171	−0.012
15 Trichloroethylene	0.403	−0.147	−0.096
16 1,2-Dichloroethylene	0.071	0.481	0.508
17 1,1-Dichloroethylene	0.007	0.500	−0.339

sion may be inhibited by the high contaminant levels which occur in the soil column at the study site.

Constructing the source signature matrix

The Q-mode factor loadings of the analytes shown in Table 3 are suggestive of the mass ratios of the contaminants in source emissions. Q-Mode factor analysis followed by Varimax rotation of the vectors is usually found to be more informative, since simple structure of the rotated factors yields a clear indication of variable correlations and the chemical nature of controls. The qualitative results have been shown to be similar to those of chemical equilibrium modeling [21]. However, the factors are abstract constructions which do not uniquely describe the data set. Their use as quantitative source signatures is arbitrary and prone to error.

A major problem with abstract factors is that they are orthogonal, whereas actual source profiles most often are not. Adequate factor representations of source profiles can be estimated with target factor analysis (TFA) and by target transformation factor analysis (TTFA). In TFA, the signatures of known sources are tested to see if they lie within factor space. The technique is described by Malinowski [22]. In TTFA, source profiles are developed through iterative development of arbitrary starting vectors. The theory and practice is described in detail by Hopke [1].

Source signatures were derived with the TTFA method using FANTASIA. The estimated concentration profiles are given in Table 4. While contaminant plumes from all sources are seen to be heavily loaded with common industrial solvents and degreasers (toluene, xylene, 1,1,1-trichloroethane, trichloroethylene, and 1,2-dichloroethylene), the source signatures are distinctive.

Due to differential chemical conversion and retardation in the soil column, the source signatures estimated with TTFA differ from those measured in groundwater samples taken near the identified sources. The source matrix can be improved by employing measured source signatures which are corrected for conversion and retardation. The corrections are made assuming first-order rate processes:

$$C_R = C_s\, e^{-kt} \qquad (2)$$

where C_R is the contaminant concentration in source material at the receptor site, C_s is the contaminant concentration in effluent at the source, k is the conversion rate estimated from receptor site data, and t is transport time estimated from the groundwater flow-rate.

Apportioning responsibility for contamination

Contamination observed at the receptor sites was apportioned among the sources using two modeling methods: TTFA and chemical mass balance (CMB). One major difference between the two is that the TTFA method employs estimated source profiles given in Table 4, while the CMB

TABLE 4

Source profile estimates derived by target transformation factor analysis (wt.% of VOC)

Analyte	Source A	Source B	Source C
4 Benzene	0.017174	0.004318	0.002961
5 Ethylbenzene	0.007345	0.031668	0.012249
6 Toluene	0.255057	0.022323	0.074446
7 Xylene	0.055741	0.077092	0.035716
8 Methylene chloride	0.044793	0.018758	0.037245
11 1,1,1-Trichloroethane	0.143002	0.002817	0.136991
12 1,2-Dichloroethane	0.023679	0.019091	0.006269
13 1,1-Dichloroethane	0.035287	0.016363	0.011702
14 Tetrachloroethylene	0.040368	0.097456	0.402743
15 Trichloroethylene	0.212233	0.035713	0.151661
16 1,2-Dichloroethylene	0.109194	0.592969	0.128001
17 1,1-Dichloroethylene	0.056128	0.081432	0.000015

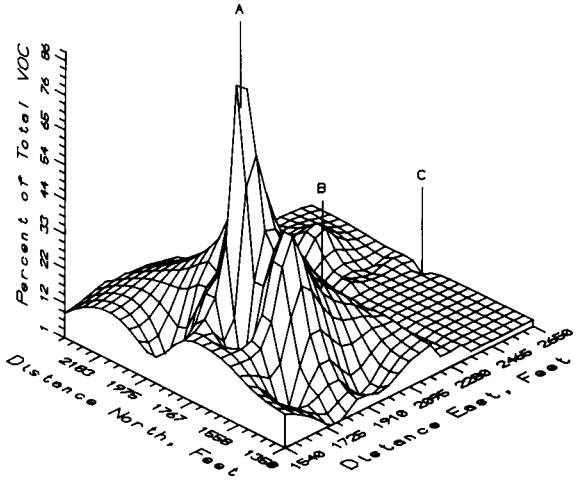


Fig. 2. CMB estimate of the percent of total VOC contributed by source A as a function of location at the study site.

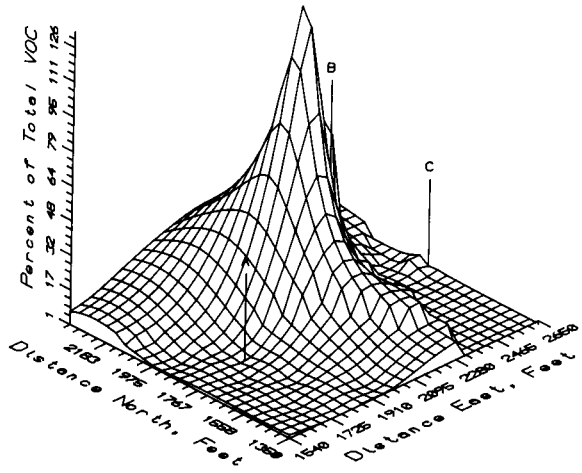


Fig. 3. CMB estimate of the percent of total VOC contributed by source B as a function of location at the study site.

method employs the source profiles measured in the field. The theory and practice of CMB has been described by Gordon [2], and Williamson and Dubose [23].

As Table 2 shows, the three-factor TTFA estimates of receptor-site contaminant levels are in error, on average, by 160%. The estimates are based on profiles that are not corrected for chemical conversion or retardation. CMB estimates of individual contaminant levels are in error by about 40%. The estimates are based on the earliest profiles measured at the identified sources. In both cases, prediction errors increase with increasing distance between source and receptor site.

The percentage of total volatile organic carbon (VOC) in receptor-site well-water samples due to sources A, B, and C are shown in Figs. 2–4, respectively. Over 60% of total VOC is accounted for in the model estimates. The two-dimensional percent composition isopleths were plotted with SURFER (Golden Software, Golden, CO) using an inverse-distance interpolation method. The figures may not be accurate near the southwest plot boundary due to lack of data. Also, few wells were sampled in the area between sources A and B. Nevertheless, the results agree with expectations based on net groundwater flow and soil column characteristics.

The receptor model estimates are biased by a number of factors. One is change in the source signatures with time due to differential chemical conversion and retardation. The receptor modeling process readily reveals the nature and rates of the physical chemical processes that are occurring in the contaminant plumes, and these can be

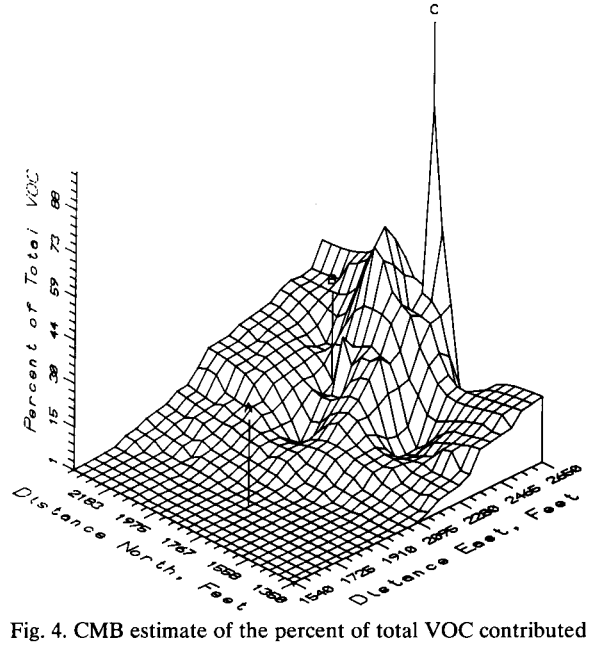


Fig. 4. CMB estimate of the percent of total VOC contributed by source C as a function of location at the study site.

accounted for by means of a non-linear model which employs corrected source signatures.

Another problem is measurement error in source signatures and receptor-site contaminant concentrations. To account for this, the model's resolution can be optimized with an effective-variance sample-weighting scheme in the least-squares regression analysis [23].

Model estimates are particularly sensitive to errors in source signatures if one or more sources do not have distinctive chemical signatures or if several sources with similar effluent composition are present. Colinearity of source signatures appears not to be a problem in the present study, but ridge regression, principal component regression, or partial-least-squares regression could be used to reduce error in the estimate if necessary [24].

Finally, all significant sources of contamination may not be included in the analysis. Obviously, omission of an important source will greatly diminish the accuracy of the prediction. There is evidence that unidentified upgradient sources contribute to groundwater contamination in the geographic area covered in this study. Poor agreement of regression estimates with data on remote wells may be due, in part, to this.

Conclusion

This report demonstrates that receptor models, widely used to study air pollution, are also applicable to the study of groundwater pollution. Chemical evolution of apparent source signatures, which is usually ignored in air pollution modeling, is a major problem that must be addressed in groundwater modeling. It results in inflated estimates of the number of pollution sources and error in source apportionment. The receptor modeling process, coupled with pattern recognition techniques, reveals the nature and rates of physical chemical processing occurring in the soil column, allowing an accurate estimate of the number of significant sources and correction of apparent source signatures. Modeling error can be reduced by use of non-linear receptor models in which corrected source signatures are used.

I wish to thank Mr. David Swanson, Mr. James Tolbert, and Dr. Richard Rediske, WW Engineering and Science, and Ms. Melissa Brown for helpful discussions and information related to this work. This work was supported, in part, by grants from the Kellogg Foundation, the Grand Rapids Foundation, and from the Water Resources Institute of Grand Valley State University.

REFERENCES

1 P.K. Hopke, Receptor Modeling in Environmental Chemistry. Wiley, New York, 1985.
2 G.E. Gordon, Environ. Sci. Technol., 22 (1988) 1132.
3 L.A. Currie, in J. Buffle and H.P. van Leeuwen (Eds.), Environmental Particles, Lewis Publ., Ann Arbor, MI, 1992, p. 3.
4 E.J. Baum and R.L. Pitter, The Nature and Origin of Atmospheric Aerosol and Visibility Reduction in Portland, Oregon, Final Report, prepared for the Oregon Department of Transportation, U.S. Department of Transportation, Federal Highway Administration, March 1976, National Technical Information Service PB-268 3T8/3WP.
5 J.A. Cooper and J.G. Watson, Portland Aerosol Characterization Study, Report to the Oregon Department of Environmental Quality, 1979.
6 J.E. Core, J.A. Cooper, P.L. Hanrahan and W.M. Cox, J. Air Pollut. Control Assoc., 32 (1982) 1142.
7 E.J. Baum, Chemom. Intell. Lab. Syst. 3 (1988) 91.
8 Draft Remedial Investigation Volume I Report, Verona Well Field, Battle Creek, MI, CH2M-Hill Inc., Corvallis, OR, 1987.
9 P.K. Hopke, D.J. Alpert and B.A. Roscoe, Comput. Chem., 7 (1983) 149.
10 I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, Berlin, 1986.
11 E.R. Malinowski and D.G. Howery, Factor Analysis in Chemistry, Wiley, New York, 1980.
12 W. Wegscheider and D.E. Leyden, Adv. X-Ray Anal., 22 (1979) 357.
13 R.R. Meglen and R.J. Sistko, in J. Breen and P.E. Robinson (Eds.), Environmental Applications of Chemometrics (Am. Chem. Soc. Symp. Ser., 292), American Chemical Society, Washington, DC, 1985 p. 13.
14 E.J. Usinoff and A. Guzman-Guzman, Ground Water, 27 (1989) 27.
15 J.W. Moore and S. Ramamoorthy, Organic Chemicals in Natural Waters, Applied Monitoring and Impact Assessment, Springer-Verlag, Berlin, 1984.
16 P.H. Howard, R.S. Boethling, W.F. Jarvis, W.M. Meylan and E.M. Michalenko, Handbook of Environmental Degradation Rates, Lewis Publ., Chelsea, MI, 1991.

17 F. Parsons, P.R. Wood and J. DeMarco, J. Am. Water Works Assoc., 76 (1984) 56.
18 F. Parsons, G.B. Lage and R. Rice, Environ. Toxicol. Chem., 4 (1985) 739.
19 P.R. Wood, R.F. Lang and I.L. Payan, in C.H. Ward, W. Giger and P.L. McCarty (Eds.), Ground Water Quality, Wiley, New York, 1985.
20 J.L. Wilson, G.B. Smith and J.F. Rees, Environ. Sci. Technol., 20 (1986) 997.
21 J.R. Kramer, Chem. Geol., 4 (1969) 37.
22 E.R. Malinowski, Factor Analysis in Chemistry, Wiley, 1980.
23 H.J. Williamson and D.A. DuBose, Receptor Model Technical Series, Vol. 3, User's Manual for Chemical Mass Balance Model, EPA-450/4-83-014, National Technical Information Service, Washington, DC, 1983.
24 I.E. Frank and B.R. Kowalski, in J. Breen and P.E. Robinson (Eds.), Environmental Applications of Chemometrics (Am. Chem. Soc. Symp. Ser., 292), American Chemical Society, Washington, DC, 1985 p. 271.

Analytica Chimica Acta, 277 (1993) 431–439
Elsevier Science Publishers B.V., Amsterdam

Optimal frequency of quality control analyses in environmental measurement

M.J. Miah, F.C. Garner, M.A. Stapanian and G.A. Laing

Lockheed Engineering and Sciences Company, 980 Kelly Johnson Dr., Las Vegas, NV 89119 (USA)

(Received 3rd September 1992; revised manuscript received 21st December 1992)

Abstract

A variety of quality control (QC) analyses are performed with regular frequency in environmental chemical measurements. Typically, an acceptable result must be obtained for each QC check before measurement of environmental samples may begin, and each check is periodically repeated to validate the analyses of intervening samples. Corrective actions are performed if the check is failed, and the intervening batch of samples are remeasured. Under such a scheme, excessive costs will occur if the quality control checks are too frequent (i.e., batch size is too small) or infrequent (i.e., batch size is too large). This paper describes an approach for investigating the frequency of QC checks for minimizing the expected cost of measuring a set of samples. Two specific expected cost models are discussed. It is shown that the empirical estimate of measurement drift can be used along with computer simulation to estimate the batch size that will minimize the expected cost of sample analyses. Measurement drift of inductively coupled plasma mass spectrometry for ^{121}Sb is demonstrated with data from standard reference materials repeatedly analyzed between QC solutions. Examples are presented using computer simulation and empirical estimates of linear drift functions for analytical measurements.

Keywords: Inductively coupled plasma mass spectrometry; Computer simulation; Quality control

Quality control (QC) analyses in environmental analytical chemistry typically consists of a series of analyses of materials of known composition. The QC may include analyses of standards in a calibration confirmation check, analyses of blank materials which are known to be void of the analyte of interest, and analyses of well-characterized test samples, often called performance evaluation samples or audit samples. Most QC is internal to the laboratory and includes well-characterized materials and associated acceptance criteria.

Correspondence to: M.A. Stapanian, Lockheed Engineering and Sciences Co., 980 Kelly Johnson Dr., Las Vegas, NV 89119 (USA).

The U.S. Environmental Protection Agency's (EPA) Superfund Contract Laboratory Program is responsible for the analyses of more than 100 000 solid and aqueous samples each year. The inclusive cost of these analyses is on the order of US $50 million annually. Each analysis is performed by one of approximately 50 pre-qualified laboratories. Unlike traditional QC objectives for a single process or plant, such a large-scale program requires that the QC protocols be established in advance by contract and be identical for each laboratory. QC materials are measured initially and periodically with the goal of ensuring the quality of environmental sample measurements [1,2]. For example, one protocol [1] requires the analysis of a blank material before and

after every group of ten environmental samples. The rationale is that the instrumental baseline response has not shifted or drifted for the ten intervening samples if the blanks are demonstrated to be uncontaminated. If the latter blank reveals contamination, then the laboratory is required under contract to take corrective actions to identify and remedy the causes of the problem. These corrective actions include (1) stop the analysis, (2) correct the problem, (3) recalibrate the instrument, (4) reverify the calibration, and (5) reanalyze the ten samples that were potentially affected by the drift. In this way, the quality of every group of samples is certifiable.

The frequency of QC checks is usually based on convenience or convention rather than on empirical evidence. It is advantageous to determine the ideal frequency of QC checks under this scheme. An unnecessarily high frequency will result in excessive expenditure on the QC analyses themselves. An unnecessarily low frequency will result in excessive expenditures for reanalyses. In this paper we attempt to minimize the expected total cost of analyzing a set of N samples, including associated QC analyses and any reanalyses that arise when the QC criteria are failed. A probabilistic model was developed by Yfantis et al. [3]. The objectives of this paper are to expand on that idea and to provide a more general methodology for establishing a priori QC criteria for multiple laboratory studies. Throughout this paper, we follow the U.S. EPA's Superfund Contract Laboratory Program protocol [1,2]. This protocol is widely used in the United States for analyses of samples from hazardous waste sites.

MODEL 1

Suppose that the results of QC analyses are independent with constant probability of success P. Success means that a measurement has passed a QC check. QC checks are performed after measuring each subset of samples we define as a batch, and batches are analyzed consecutively. We shall consider specific cases first and then the results of specific cases will be generalized. In all cases we assume that the analysis cost, C, for each sample is constant and that the cost of one QC check is equal to the cost of a single analysis.

Case 1

Let us consider the case when the QC checks are performed after N samples are analyzed. All possible events related to this case can be written as

$$E_1,\ F_1 \cap E_2,\ F_1 \cap F_2 \cap E_3, \ldots,$$
$$\cap_{i=1,n} F_i \cap E_{n+1}, \ldots$$

where F_i indicates that the measurement failed the QC check in the ith trial and E_i means that the measurement passed the QC check in the ith trial. The sum of the probabilities for all possible events can be written as

$$P(E_1) + P(F_1 \cap E_2) + P(F_1 \cap F_2 \cap E_3) + \ldots + P(\cap_{i=1,n} F_i \cap E_{n+1}) + \ldots$$

Considering the independence of the events and the constancy of the probability of the measurement for passing the QC check, the above sum can be written as

$$\sum_{1}^{\infty} P_b Q_b^{i-1} \tag{1}$$

where $i = 1, 2, \ldots, \infty$, P_b is the probability that the measurement will pass the QC check after the batch is analyzed, and $Q_b = 1 - P_b$. In this case, the batch is the entire set of N samples. Since $0 \leqslant Q_b \leqslant 1$, the sum of the series in Eqn. 1 is equal to 1. This indicates that we have considered all possible cases.

Let $C_0, C_1, \ldots, C_\infty$ represent the costs of the scenarios in which there are zero, 1, and up to infinite QC failures, respectively. The expected cost equation in this case is as follows:

$$E(C) = \sum_{1}^{\infty} C_{i-1} P_b Q_b^{i-1} \tag{2}$$

Recall that a measurement may fail the QC test in any trial. It can be shown that $C_0 = C(N+1)$, $C_1 = 2C(N+1), \ldots, C_{n-1} = nC(N+1)$. Inserting these values of C_i into Eqn. 2, the expected cost equation becomes

$$E(C) = C(N+1)/P_b$$

Case 2

Let us consider the case where there are two QC checks on the measurement system during the analysis of N samples. Assume that N is even and that $N/2 = K$ is the number of samples within each batch. The QC checks are performed after each batch of sample analysis. All possible events related to this case may be written as

$$E_1^1 \cap E_2^2,\ E_1^1 \cap F_2^2 \cap E_3^2,\ F_1^1 \cap E_2^1 \cap E_3^2, \ldots$$

where F_i^j indicates that the QC check for the jth batch failed in the ith trials and E_i^j indicates that the QC check for the jth batch passed in the ith trial. The sum of the probabilities for all possible events can be written as

$$P(E_1^1 \cap E_2^2) + P(E_1^1 \cap F_2^2 \cap E_3^2) + P(F_1^1 \cap E_2^1 \cap E_3^2) + \ldots$$

Considering the independence of the events and the constancy of the probability of the measurement passing the QC check, the above sum can be written as

$$\sum_1^\infty i P_b^2 Q_b^{i-1} \tag{3}$$

where P_b is the probability that K samples will be analyzed sequentially without a QC failure. Since $0 \leqslant Q_b \leqslant 1$, the sum of the above series is equal to 1. This indicates that we have considered all possible cases. Again, let $C_0, C_1, \ldots, C_\infty$ represent the costs of the scenarios in which there were zero, one, and up to infinite QC failures, respectively. The expected cost of analysis can be expressed as

$$E(C) = \sum_1^\infty i C_{i-1} P_b^2 Q_b^{i-1} \tag{4}$$

It can be shown that $C_0 = 2C(K+1)$, $C_1 = 3C(K+1), \ldots, C_n = (n+2)C(K+1)$. Inserting these values of C_i in Eqn. 4, the expected cost equation becomes

$$E(C) = C(K+1)P_b^2 \sum_1^\infty i(i+1)Q_b^{i-1} = 2C(K+1)/P_b \tag{5}$$

Case 3

Let us consider the case where n QC checks are to be performed on the measurement system during the analysis of N samples. Assume that

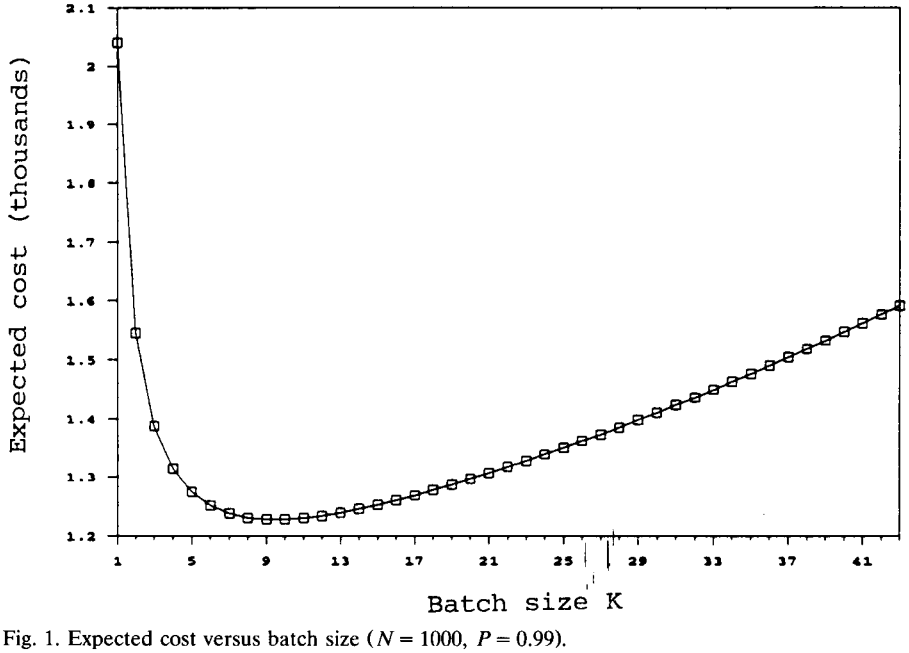


Fig. 1. Expected cost versus batch size ($N = 1000$, $P = 0.99$).

each batch is of size $N/n = K$. The QC checks are performed after each batch of sample analyses. The general equation for the probability of each event can be written as

$$P_i = \binom{n+i-1}{i} P_b^n Q_b^i \quad (6)$$

where $i = 0, 1, 2, 3, \ldots, \infty$ and where P_i represents the probability of i QC failures during the analysis of n batches of samples. The sum of the probabilities for all possible events can be written as

$$\sum_0^\infty \binom{n+i-1}{i} P_b^n Q_b^i \quad (7)$$

The events are independent and the probability that the measurements pass the QC checks is constant. From previous work [4], it can be shown that the sum in Eqn. 7 is equal to 1. This result indicates that we have considered all possible cases. As before, let $C_0, C_1, \ldots, C_\infty$ be the associated costs of respective events. The expected cost of analyses in this case can be expressed as

$$\sum_0^\infty \binom{n+i-1}{i} C_i P_b^n Q_b^i \quad (8)$$

It can be shown that $C_0 = nC(K+1)$, $C_1 = C(K+1)(n+1), \ldots, C_n = C(K+1)(2n+1)$. Using these values of C_i in Eqn. 8, the expected total cost $E(C)$ is

$$C(K+1)P_b^n \sum_0^\infty \binom{n+i}{i} Q_b^i = nC(K+1)/P_b \quad (9)$$

Since $P_b = P^{K+1}$ in this model, one needs to know the value of P in order to determine the value of K which gives the lowest expected cost. For a given C and P, different $E(C)$ can be plotted against the corresponding value of K. This plot can be used to determine the batch size that will minimize the expected total cost. Figure 1 shows that for $N = 1000$ and $P = 0.99$, optimal batch size is $K = 9$.

MODEL 2

In this model we shall no longer assume that P_b is constant. Typically, the probability that the measurement system will pass or fail a QC check in a given trial depends on the state of the measurement. We shall assume that in sequential

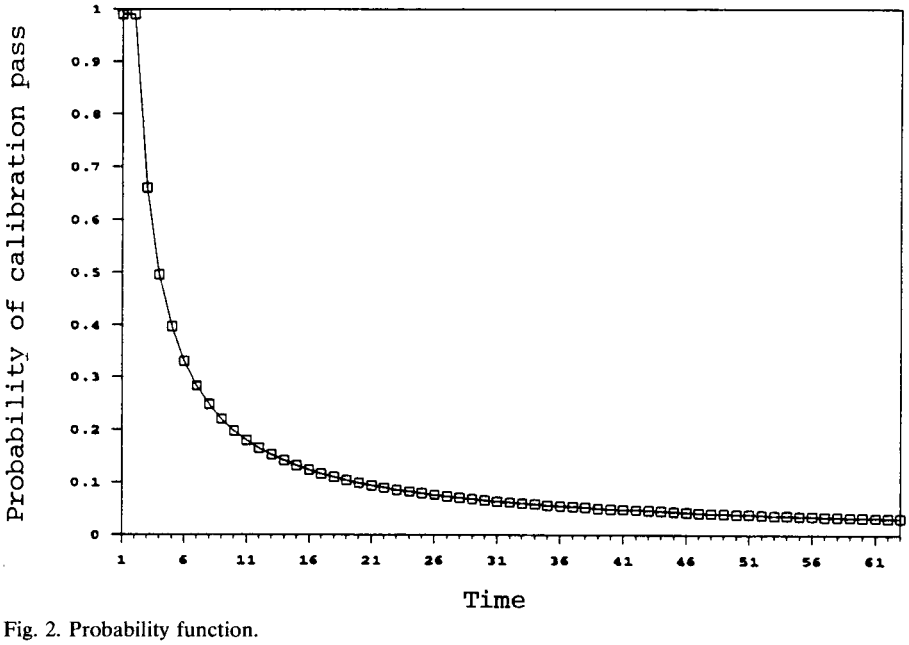


Fig. 2. Probability function.

analyses the measurement follows a drift locus. We shall also assume that the locus is a continuous curve, and that once the measurement system has failed the QC check, the subsequent corrective action will reset the system to its initial state. That means that the drift locus follows the same curve between any two sequential failures.

Let us assume that Fig. 2 represents the probability drift curve for a given measurement system. The *Y*-axis shows the conditional probability that the measurement will pass the QC check after performing the number of analyses on the *X*-axis, including samples and QC. Once corrective action is taken, the locus of the probability curve is reset and the drift curve follows the same path. Suppose that there are n batches each of K samples to be analyzed. The expected cost function has the following general form:

$$E(C) = P_0C_0 + P_1C_1 + P_2C_2 + P_3C_3 + \ldots$$

where P_i represents the probability of i QC failures in n batches of analyses and C_i is the corresponding cost of sample analysis for n batches including the QC cost of i failures. Theoretically, there could be an infinite number of QC failures, but for every $\epsilon > 0$, there exists a ∂ such that $i > \partial$ will imply $P_i < \epsilon$. Further, C_i is always finite and increases at a slower rate than the rate at which P_i tends to 0. Hence we can assume that P_iC_i is trivially small for large i.

Case 1: No QC failures

In this event, $P_0 = \Pi_{i=1,n}p_i$, where P_0 is the probability that all batches were analyzed without a QC failure, n is the number of batches, and p_i is the probability that the measurement process will pass the ith QC check.

Case 2: One QC failure

If there is exactly one QC failure and if that failure occurs in batch t, then the probability of that event is $(1-p_t)\Pi_{i=1,t-1}p_i\Pi_{j=1,n-t+1}p_j$. The second factor is $\Pi_{j=1,n-t+1}p_j$ instead of $\Pi_{j=t,n}p_j$ because once the measurement process fails the QC, corrective action resets the measurement and the drift curve. The single failure can occur in any of the n batches. Hence, $P_1 = \Sigma_{t=1,n}(1-p_t)\Pi_{i=1,t-1}p_i\Pi_{j=1,n-t+1}p_j$, where P_1 is the probability that all batches were analyzed with one QC failure, n is the number of batches, and p_i is the probability that the measurement process will pass the ith QC check.

Case 3: Two QC failures

If there are only two calibration failures and if failures occurred at the tth and the ith batch, then the probability of that event is $(1-p_t)(1-p_i)\Pi_{s=1,t-1}p_s\Pi_{j=1,i-1}p_j\Pi_{k=1,n-t-j+2}p_k$. The third and the fourth factors are $\Pi_{j=1,i-1}p_j$ and $\Pi_{k=1,n-t-j+2}p_k$, respectively, because the first failure occurred at the tth QC check and the second failure occurred at the ith QC check. The first failure can occur anywhere between the first trial to the nth trials, the second failure can occur only among last $n-t+1$ trials. Hence, the overall probability of this event of two failures during the analysis of n batches of samples is $P_2 = \Sigma_{t=1,n}\Sigma_{i=1,n-t+1}(1-p_t)(1-p_i)\Pi_{s=1,t-1}p_s\Pi_{j=1,i-1}p_j\Pi_{k=1,n-t-i+2}p_k$, where P_2 is the probability that all batches were analyzed with exactly two QC failures, and p_i is the probability that the measurement process will pass the ith QC check.

Case 4: n QC failures

Similarly, the general expression for n_1 number of failures during the analysis of n batches can be written as

$$P_{n1} = \Sigma_{(S)}\Pi_S(1-p_S)\Pi_S\{\Pi_* p_s\}$$

In this case, n_1 is assumed to be less than n. In general, S is the set of indices representing the stages of QC failures. When there are two QC failures $S = \{t, i\}$. $\Sigma_{(S)}$ is the sequential sum over the set S. For example, if $S = \{t, i, j\}$, then $\Sigma_{(S)} = \Sigma_{t=1,n}\Sigma_{i=1,n-t+1}\Sigma_{j=1,n-t-i+2}$; Π_S is the product over the set S, and $\Pi_S\{\Pi_*\} = \Pi_{s=1,t-1}\Pi_{j=1,i-1}\Pi_{k=1,j-1}\Pi_{1=1,n-t-i-j+3}$. Sometimes the instrument drift function can be estimated directly. In those cases, it may not be necessary to obtain the locus of the probability drift curve in order to estimate the sample size which minimizes the analytical cost. As shown below, the mathematical complexity of determining the optimal frequency of QC may be avoided through computer simulation and experimentation. However, such a simulation could be re-

stricted to the underlying assumptions. In order to reflect any real world situation, it is essential that the underlying assumptions of any computer simulation are consistent with the underlying principal of the real world situation.

Computer simulation

Suppose that a measurement is controlled with respect to bias through initial calibration and periodic calibration verification. Suppose further that the acceptance criteria for bias is an acceptance interval (L, U) and that the mean bias drifts linearly, the measurement response is Gaussian with constant variance. Whenever a calibration verification check is observed outside of the acceptance interval, corrective action is performed by setting the bias to zero, another check is performed, and the last batch of K samples is reanalyzed. The bias again continues to drift, and the cycle is repeated until all N samples are analyzed between acceptable calibration checks. In this example, $N = 25\,200$ and the cost of a calibration verification and the cost of sample analysis were both set equal to one unit. This process was investigated for various slopes, expressed as standard deviations (S.D.) per analysis, of the linear bias drift and batch sizes. The average cost, which is proportional to the mean number of analyses, was estimated by simulating the process on a computer 10 000 times for each slope and batch size.

As Fig. 3 illustrates, the expected cost depends on both slope and the batch size. The optimal batch size, for which expected cost is minimized, varies according to the slope. A relatively large slope of 0.10 s per analysis yields minimal cost when $K = 12$, whereas a relatively small slope of 0.02 s per analysis yields minimal expected cost when K is about 30. Substantially increased costs are expected whenever K is much greater or less than optimal. Reduced cost and flatter curves are characteristic of the more stable measurement scenarios (i.e., those having lower slope).

This example illustrates that considerable cost savings can occur when optimal K is selected from the typical slope of drift function of the measurement system. Further, costs can be reduced by modifying or selecting the measurement procedure so that the typical bias drift slope is minimal.

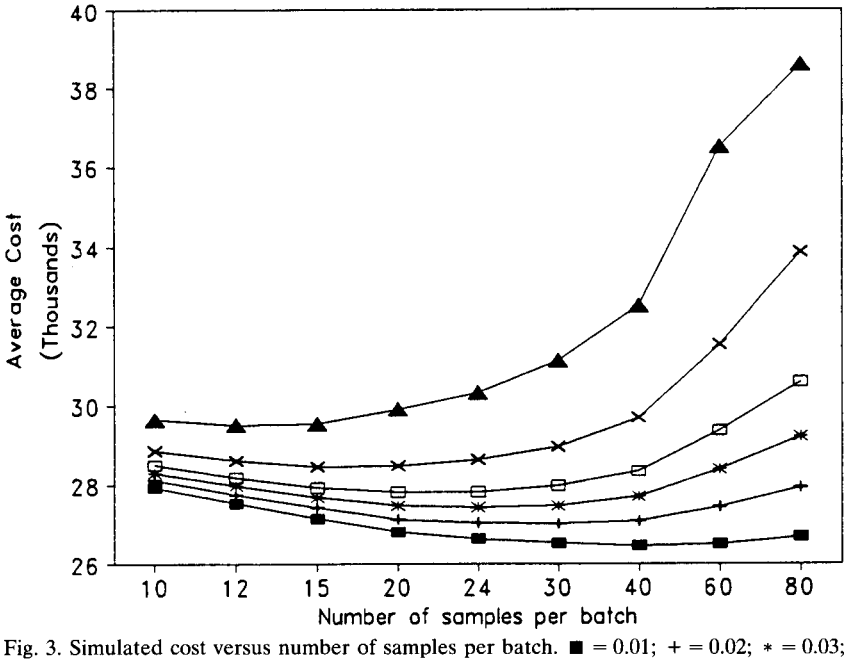


Fig. 3. Simulated cost versus number of samples per batch. ■ = 0.01; + = 0.02; * = 0.03; □ = 0.04; × = 0.06; ▲ = 0.10.

TABLE 1

The sample size corresponding to various instrument drift slopes (β)

β	0.01	0.02	0.03	0.04	0.06	0.1	0.2
K	40	30	24	20	15	12	10

Each instrument may have its own drift parameters leading to optimal batch size that will minimize expected total cost. Table 1 lists the optima corresponding to various instrument drift slopes (β). As above, $N = 25\,200$ and the number of simulations = 10 000. β is expressed in standard deviations per analysis.

Optimal batch size depends not only on the slope parameter but also on the QC analyses acceptance window (U, L). Table 1 was generated by using ± 18 times standard deviation of the instrument error as the acceptance window. The larger the window, the larger the optimal batch size will be. Similar results would be observed if the drift was in the imprecision rather than the bias, or even if both were drifting. If the information regarding the drift function is available, then the optimal batch size (and, correspondingly, optimal QC frequency) can be estimated. For many measurement systems, however, the drift information is not readily available. In those cases, empirical estimation of the drift function, such as that shown below, can be obtained by repeated measurement of standard samples.

EXPERIMENTAL

ICP-MS instrument stability

The purpose of this experiment was to evaluate measurement drift using inductively coupled plasma mass spectrometry (ICP-MS) under conditions that would degrade instrument performance. In ICP-MS, test portions are physically drawn through an orifice that leads to high vacuum. If the samples contain high concentrations of dissolved solids, the orifice may become constructed by deposition. Instrument performance may be affected as a result. This experiment repeatedly introduced samples containing high concentrations of dissolved solids during an extended analytical run.

The test portions repeatedly analyzed between QC solutions consisted of NIST Standard Reference Material (SRM) 2704 Buffalo River Sediment or NIST SRM 1645 River Sediment. In each test portion, 0.500 g of the SRM was weighed in a digestion vessel and then digested in a microwave oven. The test portion was diluted to 200 ml prior to analysis. Parameters for the ICP-MS instrument and plasma used in this experiment are listed in Table 2.

Ten SRM test portions were analyzed between each set of QC solutions. The QC solutions consisted of a continuing calibration verification (CCV) followed by a continuing calibration blank solution. The concentrations of the CCVs were monitored during the analytical run and are reported without using an internal standard to correct for instrument drift. The total analytical run was 14.93 h in duration, beginning at 09:32 and ending at 00:28 the next day. A total of 115 analyses occurred during this time, including calibration standards, QC solutions, and SRM test portions.

TABLE 2

ICP-MS and plasma parameters

ICP-MS instrument parameters	
Manufacturer and Model:	Perkin-Elmer Sciex Elan Model 250 upgraded to Model 500
Sweeps per replicate:	200
Number of replicates:	3
Points per peak:	1
Resolution:	High
Polarity:	+
CEM voltage:	3.8 kV
Plasma parameters	
Plasma gas:	12.0 l min^{-1}
Nebulizer gas:	0.99 l min^{-1}
Auxiliary gas:	1.6 l min^{-1}
Power:	1.2 kW
Nebulizer:	Meinhard type TR-30-C3

RESULTS

In order to illustrate this empirical technique for determining the optimal frequencies for QC samples, only the results for one analyte, ^{121}Sb, are presented. Although the drift curves were analyte-specific, the results for ^{121}Sb are qualitatively typical for this experiment. The limits of acceptability were set a priori at $\pm 10\%$ of the true concentration in the CCVs.

The true concentration of ^{121}Sb in the CCVs was 655 ppb. By using standard linear regression technique, the following estimated linear drift function is obtained:

$$C = 672.814 + 0.690X$$

where C = the concentration of ^{121}Sb expressed in ppb and X = the ordered sequence of analysis (i.e., the number of experimental "runs"). The standard errors were 7.349 and 0.071 for the y-intercept and slope, respectively, and $r^2 = 0.922$.

DISCUSSION

The drift function for ^{121}Sb in this study was determined by repeatedly analyzing standard reference materials (prepared from river sediments) between QC solutions. Typically, test portions in a batch may have diverse origins and the analyte-specific drift functions may be concentration dependent. In such a case, using a step-function technique [5], intervals of concentrations may be defined for each analyte and drift functions within the intervals may be determined. An empirical determination of the drift–concentration relationship is beyond the scope of this paper.

Instruments that simultaneously measure more than one analyte present the problem of which of the optimal QC frequencies to select. A conservative choice, assuming that the analytes are of equal importance, is the frequency corresponding to the analyte which drifts out of the limits of acceptability after the fewest number of experimental runs. Drift functions may be instrument- or operator-specific, nonlinear or even cyclic. Although higher-order models are more complex, these same techniques can be used to find optimal QC frequency.

Measurement behavior should be investigated and successfully characterized in order to determine optimal K and thus minimize costs. For almost any measurement system, it would be beneficial to examine the curve of expected cost versus batch size. Without such investigation, the user risks incurring excessive costs. Further research is necessary to evaluate the uncertainty risk associated with these methods, where risk is defined as the expected loss when the decision to continue analyzing samples without recalibration is incorrect.

As mentioned, the U.S. Environmental Protection Agency is responsible for the analyses of more than one hundred thousand test portions (samples) of soils, sediments, drinking waters, ground waters, and other matrices each year for various classes of toxic chemical constituents such as pesticides, volatile organic compounds, metals, radionuclides, cyanide, and chlorinated dioxins. The bulk of these measurements are performed by contract laboratories. The quality of these analyses must be assured. There must be a verifiable high probability of achieving the objectives of the study at hand. However, standard analytical measurement methods are constructed and contractually put in place long before the studies are conceived (e.g., [6]). The level of QC must therefore be prespecified and built into the measurement methodology, and thus must be generally applicable to a variety of environmental studies. Typically, the frequencies of QC samples are based on arbitrary criteria (e.g., every twenty samples). Considerable cost savings may be realized through these research programs and the resulting careful selection of QC frequencies, based on models and experiments such as those described above.

In many scenarios, the producer of a product is responsible for establishing the QC protocols and criteria. In these instances, CUSUM charts [7–9] are often a preferred technique for predicting when, on a particular experimental run, the measurement system will be out of acceptable control. These charts are often laboratory-, operator-, or experimental run-specific. Guidelines for CUSUM charts for trend detection are available. In large-scale, multiple-laboratory studies, how-

ever, QC criteria and protocols are typically established in advance by the client, not by the analytical laboratory. The researcher may find the procedures described in this paper to be more straightforward and more practical than CUSUM charts for establishing a priori QC criteria for multiple laboratory studies. These procedures establish a minimum cost frequency as an instrument or method characteristic, whereas the CUSUM chart provides detailed QC information for a particular experimental run. The approach has the inherent advantage of simplifying data review and standardizing the QC procedures among a number of contracted laboratories. When the same type of instruments and methods are used by contract laboratories, the drift functions may be used to establish QC criteria and protocols for instruments and methods across all laboratories. Further, these procedures are useful for evaluating instrument or method performance over time.

We thank an anonymous reviewer for helpful comments on the earlier version of this paper. Although the information in this paper has been funded wholly by the United States Environmental Protection Agency under the contract number 68-C0-0049 to Lockheed Engineering & Sciences Company, it has not been subjected to Agency review. It therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. The mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFERENCES

1 U.S. Environmental Protection Agency. USEPA contract laboratory program statement of work for inorganics analysis—multi-media, multi-concentration, Document Number ILM02.0, Washington, DC, 1991.
2 U.S. Environmental Protection Agency. USEPA contract laboratory program statement of work for organics analysis —multi-media, multi-concentration, Document Number OLM01.8, Washington, DC, 1991.
3 E.A. Yfantis, G.T. Flatman and F.C. Garner, Chemom. Intell. Lab. Syst., 3 (1988) 39.
4 K.L. Chung, A Course in Probability Theory, Academic Press, New York, 1974, p. 168.
5 M.J. Miah, J. Chemom., 5 (1991) 201.
6 S.K. Drouse, D.C. Hillman, L.W. Creelman and S.J. Simon, National Surface Water Survey, Eastern Lake Survey (Phase I—Synoptic Chemistry) Quality Assurance Plan, U.S. Environmental Protection Agency, EPA 600/4-86/008, 1986.
7 H.M. Taylor, Technometrics, 10 (1968) 479.
8 C.D. Lewis, Med. Biol. Eng., 9 (1971) 315.
9 D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: a Textbook, Elsevier, Amsterdam, 1988, pp. 96–97.

Analytica Chimica Acta, 277 (1993) 441–453
Elsevier Science Publishers B.V., Amsterdam

Development of an expert system for selection of experimental designs

Ramon A. Olivero

Lockheed Environmental Systems and Technologies Company, Las Vegas, NV 89119 (USA)

Sridhar Seshadri

Digital Consulting & Software Services, Sugarland, TX 77478 (USA)

Stanley N. Deming

Department of Chemistry, University of Houston, Houston, TX 77204-5641 (USA)

(Received 3rd September 1992)

Abstract

An expert system has been developed to assist chemists in the selection of experimental designs for research projects. The system (named DXPERT) ranks thirteen types of experimental designs according to their suitability for projects presented by users in an interactive session. Design categories included are factorial, response surface, sequential simplex optimization, simplex mixture, and statistical testing. A desirability index is assigned to each design alternative according to project characteristics (attributes). Characteristics are interpreted based on expert knowledge built into the system. DXPERT uses mathematical concepts to mimic features of human intuition and decision making. Expert knowledge is represented by relevance factors (a number between minus one and plus one) in a multiple-alternative multiple-attribute table. Relevance factors are interpreted as fuzzy values that represent the degree to which a design belongs to the set of suitable designs. The formula for calculating design desirabilities is based on fuzzy mathematics. For efficiency purposes, the order of the questions presented to the user is driven by a maximum potential information gain algorithm. Design desirability indexes were found to be useful to researchers for the elimination of unsuitable designs and concentration of further efforts in the most applicable designs. A validation test was conducted with the participation of four other experts in the field.

Keywords: Expert systems; Experimental design; Chemometrics; Fuzzy logic; Desirability functions; Information theory

One goal of the analytical chemist is the acquisition of chemical information together with an estimate of the reliability of that information. The use of statistical experimental designs provides researchers with tools to gather data about chemical systems in an efficient and reliable manner [1]. Formally-designed experiments (specially statistically-based designs) can reveal information about a chemical system in the smallest number of experiments and provide an estimate of the reliability of the results. Well-characterized experimental designs have been developed in the areas of chemometrics and other scientific disciplines. Standard design types specify the level

Correspondence to: R.A. Olivero, Lockheed Environmental Systems and Technologies Co., 980 Kelly Johnson Drive, Las Vegas, NV 89119 (USA).

arrangement for the variables involved, implementation approach, and data analysis procedures. Common experimental design categories include factorial, response surface, sequential simplex optimization, simplex mixture, regression, and statistical testing. Chemometrics involves "the use of statistics to design or select optimal measurement parameters" [2]. Much of today's research in chemometrics focusses on data treatment techniques since today's computing resources make it relatively easy to evaluate collected data in many ways.

Chemical researchers are frequently faced with the problem of choosing an approach for designing experiments for a research project. This decision is based on what is known about the system, what needs to be known, the available resources and conditions, and knowledge about the candidate experimental designs. While many situations in chemical research can be addressed with common design solutions, researchers without strong background in statistical design must contend with scattered literature sources. Many chemical researchers do not have knowledge of the available designs nor the expertise to discriminate among potential experimental designs. They do not always have access to experts in the field of experimental design who can recognize the characteristics of a problem that relate to design selection. As a consequence, the application of seemingly sophisticated statistical methods to the analysis of data acquired from poor experimental designs leads to undetected erroneous conclusions about the results [3].

The objective of this work is to investigate the development of a computerized tool to help researchers in the preliminary evaluation of various experimental design classes for a given project. The tool is intended to provide advice in a way similar to an expert in the field, integrating knowledge and experience. It should efficiently gather information about the project, use the information to rank the experimental designs according to suitability, and provide help to the researcher. The scope of the software is the selection of appropriate experimental design types and does not include setting specific experimental parameters once an approach has been selected. The latter is a more mechanical task that can be accomplished with the use of existing statistical software packages. The target users are expected to have low to moderate knowledge of statistics.

SYSTEM DESIGN AND DEVELOPMENT METHODOLOGY

The problem-solving ability of an expert in the domain of experimental design for chemical research was imbedded in a custom-written computer program (named DXPERT) [4]. The resulting program is an *expert system*. Expert systems are a class of computer programs that "use knowledge about and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution" [5]. These systems are usually applied to narrow and specific domains of knowledge. *Knowledge systems* are a similar type of artificial intelligence application. Expert systems differ from knowledge systems in that the former contain the knowledge of one or more human experts (as opposed to "book" knowledge) and their assessments and recommended solutions reflect the personal experience of the particular experts [6]. Application of expert (and knowledge) systems to chemistry problems is a subject of chemometric research [7].

The development of an artificial intelligence application such as DXPERT starts with the observation of the human decision-making process of interest. Then computer models and algorithms are devised to imitate the thought process itself or the results of it. Lastly, the system is tested, refined and, if successful, implemented for routine use. These systems are intended to complement human abilities, rather than substitute for them. The basic requirement for DXPERT is to gather information about the researcher's goals, resources, and constraints and use that information to recommend suitable experimental designs. This is a classification-type expert system problem. General software requirements identified for the system include ease of use, educational value (as opposed to a black-box approach), and efficient questioning of the user.

The questioning should be logical and relevant, and the solution should be pursued with a clear sense of direction, reaching it through an efficient path.

Main components of DXPERT are the knowledge base (the representation of the information acquired from human expert) and the inference engine (the driver that guides the search for a solution). There are commercially-available expert system "shells" that provide much of the framework for development. Most shells use IF–THEN rules as a knowledge representation scheme [8]. IF–THEN rules relate a situation (premise) to an action (consequence). This option was discarded for developing the current system because of the complexity of maintaining the knowledge base and the preferability of a novel, more intuitive approach.

Custom knowledge representation scheme and inference mechanism were developed to handle the experimental design selection problem. Expert systems deal primarily with symbolic, rather than numerical information. DXPERT applies mathematical principles commonly used in chemometrics, i.e., fuzzy logic and information theory, into empirical formulas in an attempt to mimic features of human intuition [9]. The software was written in the Pascal computer language for IBM PC and compatible microcomputers running under the MS-DOS operating system.

KNOWLEDGE REPRESENTATION

The components of a simple model for an analytical chemical system are inputs (variables), process, and outputs (responses). Example of inputs are sample characteristics, analytical conditions, amount of reagents, laboratory conditions, technicians, time of analysis, etc. Most variables (or factors) for chemical systems are quantitative, with a continuum of levels.

The main research goals in analytical chemistry are optimization of responses and understanding of the underlying process. Frequently encountered project objectives include maximization of sensitivity, minimization of limit of detection, increase of accuracy and precision, assessment of reproducibility, reduction of cost, identification of important variables, quantification of variable effects, mathematical modeling, ruggedness testing, method validation, and mixture formulation.

Research objectives, analytical system characteristics, resources, constraints, and researcher preferences define the profile of a project. For the purpose of computerization, a project profile can be represented as a set of values for a number of attributes. Attributes can be nominal (categorical) or linear (numerical). For example, "optimization" could be the value for the nominal attribute PROJECT OBJECTIVE and "3" could be the value for the linear attribute NUMBER OF VARIABLES.

At the knowledge base level, attributes in DXPERT can have binary values only, namely "true" or "false." This is done to equalize the attributes so they can be treated uniformly. Complex nominal attributes, such as PROJECT OBJECTIVE, are decomposed into several binary-valued attributes: OBJECTIVE IS OPTIMIZATION, OBJECTIVE IS MODELING, OBJECTIVE IS SCREENING, etc., one of them being "true" and the rest "false." Attributes with numerical values are decomposed in ranges or individual values.

Expert knowledge in DXPERT is represented in a multiple-alternative multiple-attribute table [10]. Each column in the table represents an experimental design *alternative* (e.g., factorial design) while each row represents an *attribute* of the project (e.g., VARIABLES ARE CONTINUOUS). The knowledge of the domain expert [6] is expressed in the form of relevance factors (ranging between minus one and plus one) at each cell in the table. A relevance factor represents the expert's opinion of the suitability of the corresponding design given a project with the corresponding attribute value. A positive relevance factor (r^+) means that the value accumulates evidence in favor of the alternative, while a negative number (r^-) means that the value accumulates evidence against the alternative. A zero entry means that the value of the attribute is neutral to the given alternative (or that the information is unavailable). Table 1 is a partial multi-

TABLE 1

Example attribute/design relevance factors table (partial)
(The top end for each attribute corresponds to its "true" value, and the bottom one corresponds to its "false" value)

Designs attributes	Means comparison	Full factorial design	Central composite	Fixed size simplex
1 variable	1.00	0.50	0.00	0.00
	0.00	0.00	0.00	0.00
2 variables	0.75	1.00	1.00	0.75
	0.00	0.00	0.00	0.00
3 variables	0.50	1.00	1.00	1.00
	0.00	0.00	0.00	0.00
4 variables	0.25	1.00	0.95	1.00
	0.00	0.00	0.00	0.00
5 variables	0.00	0.90	0.90	0.95
	0.00	0.00	0.00	0.00
Optimization	0.00	0.00	0.90	1.00
	0.00	0.00	0.00	−0.10
Modelling	0.00	0.30	1.00	0.50
	0.00	0.00	−0.10	0.00
Response surface	0.00	0.60	1.00	1.00
	0.00	0.00	−0.10	−0.10
Screening	0.00	0.50	0.00	0.00
	0.00	0.00	0.00	0.00
Method comparison	1.00	0.50	0.00	0.00
	0.00	0.00	0.00	0.00
Factor compensation	0.00	0.80	1.00	0.70
	0.00	0.00	0.00	0.00
Noisy system	0.00	1.00	1.00	−0.20
	0.00	0.00	0.00	0.50
Sequential experiments	0.00	0.00	0.00	0.50
	0.00	0.00	0.00	0.00

ple-alternative multiple-attribute table showing relevance factors (both r^+ and r^-) for several attributes of a few candidate experimental designs.

The impact of the affirmation of the attribute ("true" value) is considered separate from the negation ("false" value). Therefore, two separate relevance factors are assigned per attribute/alternative combination. To illustrate the need for this differentiation, consider the case of the TIME DRIFT attribute: the fact that there is time drift in a system could make sequential-type experimental designs less desirable, but the fact that there is no time drift does not necessarily makes them desirable. In this case, a "true" value for TIME DRIFT would have a strong negative relevance factor for sequential type designs (e.g., sequential simplex) while a "false" value could get a neutral relevance factor.

Relevance factors are subjective and can be fine-tuned to increase the accuracy of the expert system. Initial knowledge acquisition for setting the relevance factors was performed by the knowledge engineer [6] through interviews with the domain expert and literature research. During DXPERT development, the domain expert became comfortable enough with the process to adjust the relevance factors to improve the expert system's performance.

The relevance factors serve two purposes in DXPERT. They are used to calculate a ranking for each candidate experimental design and to direct the order in which questions are asked from the user.

EXPERIMENTAL DESIGNS SELECTION

DXPERT includes thirteen common types of experimental designs for selection. Full factorial, fractional factorial, Box–Behnken, and central composite designs fall into the general category of response-surface designs; these are suitable for system modeling and for understanding the effects of the different variables. Saturated fractional factorial and Plackett–Burman designs are useful for screening significant variables. Sequential simplex designs (fixed and variable size) are powerful techniques for system optimization. Latin squares, a block-type design, can estimate the effect of one variable while minimizing the effects of two interfering variables. Calibration designs relate instrument response to analyte concentration with a measure of confidence on the results. Means comparison and randomized paired comparisons are useful for statistical inference about the significance of the effect of changes in experimental conditions. Simplex mixture designs are applicable for evaluation of formulations and determination of the effects of the different constituents.

Each design in DXPERT receives an overall *desirability* value, D, after each project attribute has been determined. A desirability (a number between zero and plus one, or between 0 and 100%) is a measure of how suitable a design is for the project described by the user [11]. Desirabilities are computed from the relevance factors by a combination of the evidence in favor of a design and against the design, given by the corresponding r^+ and r^- values, respectively. Zero relevance factors have no effect. R_i^+ is the cumulative evidence for a design i, while R_i^- is the cumulative evidence against the design. These terms are computed in DXPERT from the r_{ij} values across all attributes j (for $j = 1, 2, \ldots m$) by the following formulae:

$$R_i^+ = 1 - \prod_{j=1}^{m} (1 - dr_{ij}^+) \tag{1a}$$

$$R_i^- = \prod_{j=1}^{m} (1 + r_{ij}^-) \tag{1b}$$

These formulae are consistent with the interpretation of relevance factors and desirabilities as fuzzy-set memberships [12].

For the positive evidence, each additional r^+ term takes R^+ closer to 1. This mimics the way favorable information accumulates to form a strong case for the use of a design. A dampening factor d is added to prevent a single r^+ from making R^+ too high.

For the negative evidence, R^- has a minimum value of zero when at least one of the r^- values is equal to minus one. No dampening factor is used for R^- because any piece of negative information can completely eliminate the possibility of using a particular experimental design in a given situation, no matter how well it does otherwise.

The overall desirability [11] for a design is calculated from R^+ and R^- by the formula:

$$D_i = R_i^+ R_i^- \tag{2}$$

R^+ makes D range from zero to 1, while R^- reduces this value according to the magnitude of the negative information. To obtain the final value of D, it is normalized against its maximum possible value and expressed as a percent (e.g., 0.5 becomes 50%). This maximum value is calculated for the case where all meaningful attributes have the right "true" and "false" values to make the design choice ideal.

As a fuzzy-set membership, the extreme desirabilities zero and one (100%), represent the respective situations where a design is totally unsuitable and where the design is perfectly suitable for the project. Intermediate values signify the degree of membership of a particular design on the set of suitable designs; that is, how *possible* (and not how probable) it is that a design is suitable for the problem at hand.

INFERENCE MECHANISM

The purpose of DXPERT's inference engine is to gather *all* the relevant information from the user about a project with a minimum of irrelevant questions asked. Values for the project attributes are provided by the user by responding to a series of questions presented by DXPERT. The basic question cycle is to select a question (attribute),

get the answer, recalculate the design rankings, and go back to select the next question.

Attributes with very similar relevance factors across all designs will contribute little to discriminate among the designs. Conversely, those attributes with very dissimilar relevance factors across designs will cause a large spread in the rankings. At each cycle, the attribute (not yet valued) with the potential to cause the largest spread in the design rankings would yield the most information. This is a variation of the problem of optimizing a classification process using information theory.

The sequence of questioning in DXPERT is directed by an algorithm based on the concept of information contents gain. At the simplest level, each attribute can be thought of as an *operator* that when applied to the set of candidate designs, X, will classify the *elements* of that set into N subsets, x_k ($k = 1, 2, \ldots, N$). The classification is based on the applicability of each design for the user's project. Information content, first proposed by Shannon [13], is a measurement of uncertainty in predicting an outcome, as in a classification event [14]. The information contents, H, from the application of an operator is given, in relative terms, by a formula of the type:

$$H = -\sum_{k=1}^{N} p_k \log_2 p_k \tag{3}$$

where p_k is a measure of the probability of an element being classified into subset k. This formula is consistent with the intuitive concept of maximum uncertainty when all p_k values are the same and no uncertainty when one of the p_k values equals 1 (and the rest equal zero). A variation of this formalism is applied in DXPERT to calculate a measure of the potential information gain from determining the value of an attribute.

For the purpose of discussion, assume two subsets, x_1, and x_2, of suitable and unsuitable designs, respectively. Let Q_j^+ be the measure of the aggregate potential of an attribute j to increase the overall design's desirabilities, and therefore augment membership in x_1. Conversely, Q_j^- is the aggregate potential to decrease the desirabilities and is related to x_2. Those quantities are calculated by DXPERT in terms of the attribute's relevance factors, according to Eqns. 5a and 5b. The preferred attribute is that one that would have the largest information contents gain between the two subsets. The measure of information contents gain is the difference between information contents for two events. In terms of our modified formalism, the information gain, G_j, from an attribute is represented by:

$$G_j = \log_2 \frac{Q_j^+}{Q_j^-} \tag{4}$$

Maximum information gain is obtained when the attribute causes the most even spread in desirabilities. This is the case when the ratio of Q_j^+ and Q_j^- is closest to 1 or G_j is closest to zero.

The quantities Q_j^+ and Q_j^- are calculated from the relevance factors r_{ij}, weighed by the corresponding R_i^- for each design, i.

$$Q_j^+ = \sum_{i=1}^{m} r_{ij}^+ R_i^- \tag{5a}$$

$$Q_j^- = \sum_{i=1}^{m} r_{ij}^- R_i^- \tag{5b}$$

r_{ij}^+ and r_{ij}^- are positive and negative relevance factors, respectively. The weighing terms give more weight to designs with a better chance of having large changes in their desirabilities. As defined before, R_i^-, is a measure of the cumulative evidence against a design. In the calculation of desirabilities, R_i^- is multiplied by the cumulative positive evidence (R_i^+), hence, the larger the value of R_i^-, the better the possibility the design has for large increments on its desirability. That feature of the formula mimics the natural tendency of an expert to pursue the solutions which appear to have the better possibilities of being validated until they are proven or until new evidence discards them.

Two G_j values can be calculated for each attribute. Each of them represent the contribution of one of the possible values assigned by the user, namely "true" and "false." There are separate relevance factors for each option. The probability for the user's choice is not known, therefore equal weight is given to each option. Two values

are calculated, G_{Tj} and G_{Fj}, and combined in the formula:

$$I_j = \left[\frac{1+G_{Tj}}{Q_j^+ + Q_j^-}\right]_T + \left[\frac{1+G_{Fj}}{Q_j^+ + Q_j^-}\right]_F \quad (6)$$

I_j is an inverse measure of the benefit from knowing the value of a given attribute. At each question cycle, the attribute with the smallest I_j will be asked. This formula is tailored to the characteristics of the problem. The left term in Eqn. 6 is the contribution from the affirmation of the attribute (T = "true"), while the right term in Eqn. 6 is the contribution from its negation (F = "false"). The numerator of each term is a measure of how even a spread the attribute could produce among the design desirabilities. A notion of information theoretical classification is that the information gain is greatest when the partition is most even. The natural logarithm term becomes zero when the ratio of the Q terms is equal to 1. In that case the numerator takes the value 1, minimizing its contribution to I_j. The denominator of each term accounts for the fact that not all attributes are relevant to all designs; it is inefficient to rate high an attribute with a very even spread but that affects only very few designs. In general, the denominator terms diminish the magnitude of I_j for attributes that affect more designs and penalizes attributes with zero relevance factors.

A time-efficient inference engine should include problem reduction capabilities, a process called pruning [6]. Attributes that become irrelevant are eliminated from consideration. Attributes are considered irrelevant when the designs to which they pertain have fallen below a preset threshold of desirability. When all attributes have been either asked or pruned the session is complete and the current desirabilities represent DXPERT recommendation for the project at hand.

TESTING AND VALIDATION METHODOLOGY

The evaluation of DXPERT was conducted by comparing various performance features against target performance criteria. Main phases of the evaluation were efficiency testing, internal evaluation, knowledge validation, and field testing. Efficiency testing covered areas such as speed, ease of use, error trapping, and appropriateness of the hardware and software environment. Internal evaluation was done by the developers. At this phase the knowledge engineer tested for inconsistencies in the encoded knowledge. The domain expert tested for errors of omission, errors in the supplied information, errors of interpretation, and accuracy of output. Pilot users evaluated DXPERT prototypes for ambiguities in the text, fatal software errors, adequacy of on-line help, repetition, flexibility, and overall impression.

Knowledge validation was carried out through a comparative study with the participation of experts in the field of experimental design. The objective was to compare the recommendations made by DXPERT for a number of test cases with the recommendations that would be typically given by experts for the same cases. Twenty test cases were designed and sent to five recognized experts in statistical experimental design (including the domain expert), all of them authors or co-authors of books on the subject of experimental design for chemical research or quality control. This evaluation group was composed of three analytical chemists and two statisticians. The experts were asked to assign desirabilities according to their expert judgement to each of the thirteen experimental designs for each case in a format similar to DXPERT's. Thirteen of the twenty cases were designed in a way that their profile would be ideal for the application of one of the thirteen design types in DXPERT; this intended to test the ability of DXPERT to identify the appropriate situations on which to recommend each design. The other seven test cases described less straight forward situations, with a higher probability for disagreement. The ordering of test cases was randomized to avoid any bias on the part of the experts due to evident trends.

Field testing was conducted with the final DXPERT prototype. Chemical researchers were asked to use the program and provide feedback on its value as a research tool and on its user friendliness. The internal evaluation and field

testing were patterned after a method described by Wade [15] for validation of the MAX system (Moisture Analysis eXpert system).

RESULTS

The software proved to be genuinely user friendly during the field testing. The user interface is menu-driven and has a built-in windowing system. The program provides basic on-screen information about the different experimental designs included in the expert system. The help system uses hypertext to allow users to request information on unfamiliar terms. A backtracking capability allows users to change their previous answers and continue through a different logical path.

The results of two example DXPERT sessions and the results of the knowledge validation study are described here.

Two example DXPERT sessions

In the first example session, a problem was presented in which the researcher is interested in modelling the factor-response relationship for

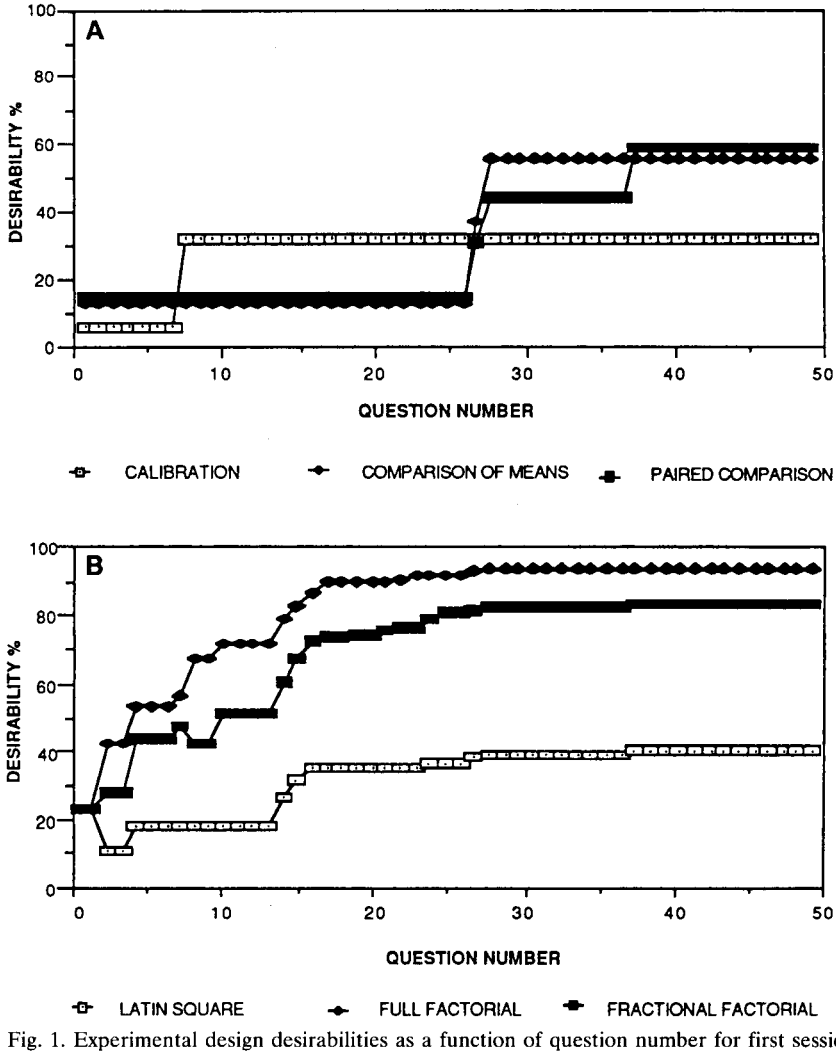


Fig. 1. Experimental design desirabilities as a function of question number for first session.

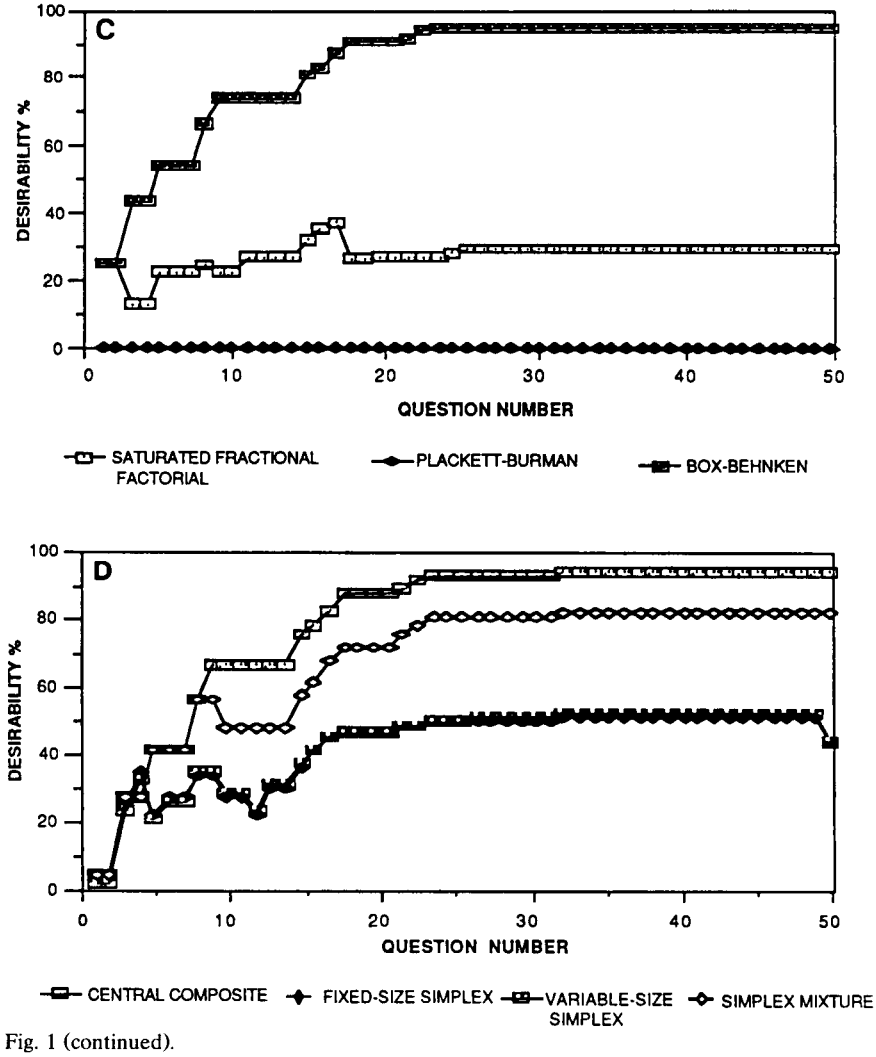


Fig. 1 (continued).

three continuous factors and one response variable. Factor interactions are assumed to be present. There is no interest in optimization of the response variable.

Figure 1 shows the progress of this first example session plotted as desirabilities vs. question number for each of the thirteen candidate experimental designs. The final recommendations of DXPERT were central composite, full-factorial, and Box–Behnken designs at 96, 95 and 95% desirability indexes, respectively (the maximum desirability is 100%). Factorial-type designs seemed to be preferred: however, the fractional factorial design reached a desirability of 85%, and mixture designs (which can be considered to be a special type of factorial design) reached 86%. Highly-fractional factorials did not seem to satisfy the project requirements: Plackett–Burman designs stayed at the zero desirability level, and saturated fractional factorial's score dropped to 27% (from 39%) when the user stated an interest in factor compensation.

Analysis of the progress of desirabilities as specific questions were answered showed that factorial-type designs followed each other's profiles, with an offset, from early in the session.

TABLE 2

Ratings by DXPERT and 5 human experts for 20 test cases involving 13 candidate experimental designs [a]

Design	D	1	2	3	4	5	D	1	2	3	4	5	D	1	2	3	4	5	D	1	2	3	4	5
CAL	14												33						33					
MEAN	14								15				14						14					
PAIR	52						27		15				38						38					
LASQ	19												18						18					
FULL	94	100	95	100	60	10	41		10				96	90	95	50		100	96	60	100	30		50
FRAC	77	90	60	40			79		80		10	100	80	30	90				80	50				
SATU	19		50					95	90	100	10	75	19		25				19					
P-B			50					100	100	10	80	75			25									
B-B	95	70	90									25	98	90	50	10			98	90	96	50		
CCD	95	70	100		40	50							98	90	90	20		50	98	90	100	100	100	10
FSPX	75	40											98	80	80				98	30				
VSPX	76	40											98	80	95				98	20				
MIX	84		90										98	100	100	100	100	75	98	90	25			
CAL	7					50							31						4					
MEAN	45		100	10	50	50							7						7					
PAIR	67	80	100			50	27						32						32					
LASQ	84	70	100	90		10	73						18						71					
FULL	92	100		10	50		36	40					95	50	100	10		90	79	10				
FRAC	80	90	70				92	70	90	90	50	100	79	40				25	78	5				
SATU			70					100	100	100	20	75	18					25	72					
P-B			70					80		10	30	75	5					25	21					
B-B												75	97	100	100	40		25	75	30	60			
CCD													97	100	100	100	100	100	75	30	60			
FSPX								30					86	30	50				96	95	90	90	40	90
VSPX								30					87	30	50				96	100	100		60	10
MIX								30					96		50				77	30	90			
CAL	28	5					7		60				28						33					
MEAN	72	80	90		40	50	65	60	80		95	50	64	100	100	100	50	100	33					
PAIR	82	100	100	100	60	75	78	90	100		5	50	65	90		20	20	75	54					
LASQ						10	83	100	100	100		100						25	83					
FULL	90		70				89	30	60				87		100				35	80		100	40	10
FRAC							89	40											92	100	100	20	60	

SATU								10												100		80			
P-B								10																
B-B																								25
CCD																								25
FSPX																								
VSPX																								
MIX																								
CAL	84	100	100	100	100	10	4					4						33						
MEAN	28						7					7						14						
PAIR	46						32					32						16						
LASQ							72					18						17						
FULL	35		90				81	5				95	30		20	40	90	94	70	85	50	30	90	
FRAC							79					78					25	76	10					
SATU							74					18					75	19					25	
P-B							21					5					75						25	
B-B							76	10	70			96	100	95	40		75	96	80	60	70		25	
CCD	68						77	10	70		20	96	100	100	100	60	100	96	80	95	40	70	10	
FSPX	84						83	100	90	10	40	90	96	60				95	50	60	10			
VSPX	85						59		100	90	40	100	96	100				96	75	90				
MIX							78	10					95	90	95			95		85				
CAL	28						33						33					38						
MEAN							14						29					21						50
PAIR	27						38						66					57						50
LASQ							17						21							70				10
FULL	37			10			94	95	50	70	60	100	96	90	90		60	100	96	90	80		100	25
FRAC	88	85	85	40	30	10	84			10	30		81											
SATU	91	85	90		10	90	21				10		20											
P-B	90	90	95	10	60	90																		
B-B	79			80		90		55	95				97	90	80		30							
CCD	81			100				55	95				97	90	85		10	50	96	85	60			
FSPX	91	60					65	10					68	10					66					
VSPX	93	95					52	35					68	10					66					
MIX	90						94		50				96						86					

[a] CAL = Calibration; MEAN = comparison of means; PAIR = paired comparison; LASQ = Latin square; FULL = full factorial; FRAC = fractional factorial; SATU = saturated fractional factorial; P-B = Plackett–Burman; B-B = Box–Behnken; CCD = central composite; FSPX = fixed-size sequential simplex; VSPX = variable-size sequential simplex; MIX = mixture designs.

This suggests that the more discriminating questions were asked at the beginning of the session and the remaining questions provided confirmation and detail checking. In general, session progress showed that related designs followed similar profiles (e.g., the central composite and Box–Behnken response–surface designs). Means comparison and randomized paired comparison designs reached a 60% desirability index, primarily resulting from questions relating to the researcher's need to compare several averages and the ability to conduct experiments in batches. Calibration designs reached only a 33% score; the only aspect of the problem that proved favorable to the use of calibration was that a continuous model was to be fitted to the data. Latin squares rose to 42%, but the inability to detect interactions with this type of design ultimately worked against it. Sequential simplexes (both fixed- and variable-size) obtained a final desirability of 56% after dropping slightly when the user stated that the experimental system drifted with time (a contraindication for the application of sequential simplex). The general effect on desirabilities of the different project attributes appeared to be consistent with the line of reasoning of an expert in the field.

In the second example session, a problem containing contrasting features was described to DXPERT for testing purposes. In an interlaboratory testing project, there is an interest in modeling the factor–response relationship for a one-variable one-response system, but the primary variable has discrete levels. This constitutes an almost impossible, worst-case situation in which the fundamental characteristics of the experimental system are incompatible with the interests of the user.

DXPERT's final desirabilities for the second session showed that none of the designs obtained a high score, which suggests that none of them is appropriate for solving the ambiguous problem presented. The highest ranking designs turned out to be randomized paired comparisons and means comparison, with 68 and 67% desirabilities, respectively. This was interpreted as meaning that the researcher was probably limited in this situation to running experiments at the different levels and determining the statistical significance of the differences observed in the responses. The calibration design was the second choice with a 49% score; were it not for the discreetness of the variable, this problem would be very suitable for a calibration design. DXPERT asked only 19 questions in the second session, as opposed to 50 in the first session. Because of the contradictory nature of the problem in the second session, a large amount of pruning was performed.

Knowledge validation study

Table 2 shows the results of a validation study involving 20 test cases submitted to DXPERT and to 5 human experts, one of whom was the original developmental domain expert.

From the desirabilities assigned by DXPERT to the test cases, it is observed that DXPERT did not select just one design, but assigned some degree of desirability to all designs that would have any feature favorable to any of the attributes of the problem. Human experts, in contrast, are seen to choose one or two designs for most cases and assign a zero desirability to the rest of the designs. (Note that although the instructions were to rate the designs on a scale from 0 to 100%, expert 4 appears to have normalized the ratings such that they add up to 100% for all designs.)

The experts did not agree in some cases, but usually there was a consensus about which was the best design for each case. For most cases the design preferred by the majority of the experts was among the top few designs selected by DXPERT. The designs to which DXPERT assigned desirabilities under 40% were usually given a zero score by the majority of the experts. In many instances, though, it was observed that DXPERT assigned high desirabilities to designs that were not considered by the experts. In cases where the experts were divided, DXPERT's advice includes the different opinions.

DISCUSSION AND CONCLUSIONS

DXPERT has demonstrated the feasibility of a microcomputer-based expert system to assist in

the selection of appropriate experimental designs. The expert system was found capable of questioning a researcher about the characteristics of an experimental problem and making appropriate recommendations concerning the suitability of each of the experimental design types included.

The validation study showed that the participating experts did not generally agree among themselves, but DXPERT followed reasonably well the specific recommendations of the domain expert that participated in the development. DXPERT achieved a degree of success in mimicking the intuition, problem solving, and decision making process of the expert.

As additional benefits, most relevant information about the problem at hand will have been gathered by the time a DXPERT session is completed. Descriptions, definitions, and references pertaining to the designs are provided in the system.

In the future, several enhancements can be implemented to improve the performance of DXPERT. The binary-value scheme for attributes would allow the flexibility of having more than one nominal value be true for a given attribute, but this was not used in the current system. The information theoretical approach to classification of crisp set members can be formally extended to classification of fuzzy set members. The knowledge representation, inference mechanism, and goal ranking approaches used in this system can be extended and applied to other suitable decision support problems in chemistry and other fields.

The authors thank the anonymous statistical experimental design experts who assisted in the evaluation of DXPERT. Although this paper has been funded wholly or in part by the United States Environmental Protection Agency under contract number 68-C0-0049 to Lockheed Engineering & Sciences Company and cooperative agreement number CR-817552-02-0 with the University of Houston, it has not been subjected to Agency review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. Mention of trade names or commercial products do not constitute Agency endorsement or recommendation for use.

REFERENCES

1 S.N. Deming and S.L. Morgan, Experimental Design: A Chemometric Approach, Elsevier, Amsterdam, 1987.
2 L.S. Ramos, K.R. Beebe, W.P. Carey, M. Sanchez, B.C. Eugenio, B.E. Wilson, L.E. Wangen and B.R. Kowalski, Anal. Chem., 56 (1986) 294R.
3 L.A. Currie, J.J. Filliben and J.R. DeVoe, Anal. Chem., 58 (1986) 2814.
4 R.A. Olivero, Selection of Experimental Designs for Analytical Chemistry with the Aid of an Expert System, Doctoral Dissertation, University of Houston, Houston, TX, 1987.
5 R. Harmon and D. King, Expert Systems, Wiley, New York, 1985.
6 R. Smith (Ed.), The Facts on File Dictionary of Artificial Intelligence, Facts On File, New York, 1989.
7 P.B. Ayscough (Ed.), Preface to Proceedings of the Meeting on Expert Systems in Chemistry and Chemical Industry, Chemom. Intell. Lab. Syst., 5 (1988) 9.
8 L. Buydens, A. Peeters and D.L. Massart, Chemom. Intell. Lab. Syst., 5 (1988) 73.
9 S. Seshadri, An Expert System Shell Based on Fuzzy Information Theory, Masters Thesis, University of Houston, Houston, TX, 1986.
10 K.R. MacCrimmon, in J.L. Cochrane and M. Zeleny (Eds.), Multiple Criteria Decision Making, University of South Carolina Press, Columbia, SC, 1973, pp. 18–44.
11 E.C. Harrington Jr., Industrial Quality Control, 21 (1965) 494.
12 L.A. Zadeh, Information and Control, 8 (1965) 338.
13 C.E. Shannon, The Bell System Technical Journal, 28 (1948) 379.
14 K. Eckschlager and V. Stepanek, Analytical Measurements and Information, Wiley, New York, 1985.
15 A.P. Wade, Modern Mathematical Methods in Analytical Chemistry, Doctoral Dissertation, University of Wales, Cardiff, 1985.

Analytica Chimica Acta, 277 (1993) 455–465
Elsevier Science Publishers B.V., Amsterdam

Designs for mixture and process variables applied in tablet formulations

C.A.A. Duineveld, A.K. Smilde and D.A. Doornbos

Research Group Chemometrics, University Centre for Pharmacy, University of Groningen, A. Deusinglaan 2, 9713 AW Groningen (Netherlands)

(Received 3rd September 1992)

Abstract

Although there are several methods for the construction of a design for process variables and mixture variables, there are not very many methods which are suitable to combine mixture and process variables in one design. Some of the methods which are feasible will be shown. These methods will be compared in a design space consisting of three mixture variables and two process variables. This comparison will be made with criteria from the theory of optimal design and with the aid of experimental data.

Keywords: Optimization methods; Design construction; Mixture variables; Pharmaceuticals; Process variables

In many real-life situations we are confronted with a combination of mixture and process variables which influence the response of systems under survey. Examples can easily be found in liquid chromatography (LC), tablet formulations but also in the construction of fish-patties. In LC the temperature, flow as well as eluent composition influence the retention times of the components. In a directly compressed tablet the composition of the tablet as well as the mixing time and the compression load influence the properties of the tablet constructed [1]. Another example is the construction of fish-patties. Here the composition, a mixture of three fish species, and the process variables, cooking temperature, cooking time and deep-fat frying time influence the final product [2].

Correspondence to: C.A.A. Duineveld, Research Group Chemometrics, University Centre for Pharmacy, University of Groningen, A. Deusinglaan 2, 9713 AW Groningen (Netherlands).

The number of designs available for the research on this kind of systems is not very large. This is because of two inherent problems associated with this problem. The first is that the mixture variables and the process variables have different characteristics. Specifically, mixture variables sum to one, whereas this does not hold for process variables. Stated otherwise, process variables can freely be varied whereas mixture variables cannot. A straightforward combination of mixture variables and process variables designs would be possible but for the second problem the number of design points tends to be very large. Therefore a scheme is necessary that limits the number of design points.

Several approaches are feasible to limit the number of design points in a mixture–process variables design. Some of these methods are introduced here. The methods applied in this paper will be explained more extensively in the theoretical section of the paper. In one of the methods first a mixture design is constructed, after which

at these mixture settings fractions of a process variables design are used. This method, which results in what we will call the fractionated design, is developed by Cornell and Gorman [3]. Another method is applicable if there are three mixture variables. Then a special four-point mixture design can be used at the settings of the process variables [4]. In a third strategy a fractional factorial design for the total number of variables (mixture variables as well as process variables) is used and a projection is used to move the infeasible settings of the mixture variables to feasible settings [5]. This is called the projection design. We will use the fractionated design as well as the projection design to construct a suitable design.

A limited number of variables, two process variables and three mixture variables, is used in the construction of the designs which will be compared. The designs are then compared to each other with criteria from the optimal design theory as well as on their effectiveness on a real data set. The real data set is a specific problem, which implies that the conclusions are only valid for the situation under which we used the designs.

DESIGN CONSTRUCTION

Fractionated design

To understand the fractionated design in mixture and process variables we start with a design in the mixture variables (in our case a seven-point mixture design). Each of these design points is used as mixture setting at which a design for the process variables is build. This is usually a fractional factorial design. The reduction in design points depends on the fractionation scheme. This strategy is shown in Fig. 1 for three process variables and three mixture components. In our example, with two process variables, this involves half fractions of a 2^2 design. The resulting design is shown in Fig. 2.

Projection design

The construction of mixture designs using projection of factorial designs has been proposed by Hau [5]. The method is based on the fact that a figure in a higher dimensional space can always be projected on a lower dimensional subspace. In our case a factorial design is constructed in the $m+p$ dimensional space (m mixture variables, p process variables), after which it is orthogonal projected onto the mixture plane, a sub-space with restriction $\Sigma\, m_i = 1$ (m_i is fraction compo-

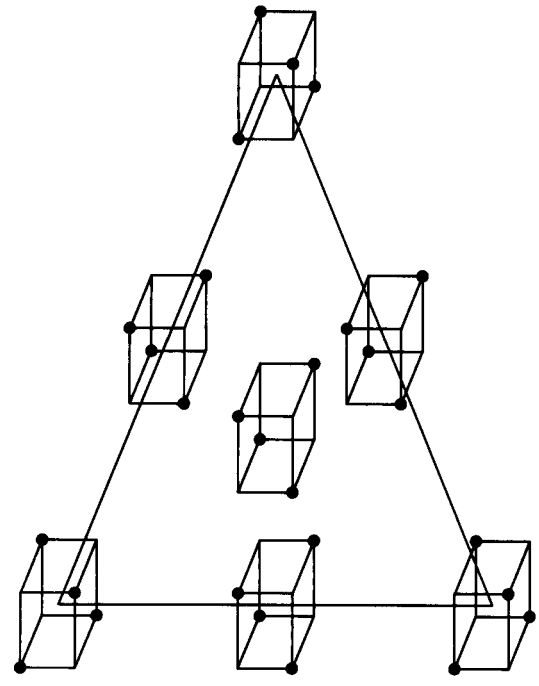


Fig. 1. Fractionated design for three process variables and three mixture variables. At the design points in the mixture design 2^{3-1} designs in the process variables are used.

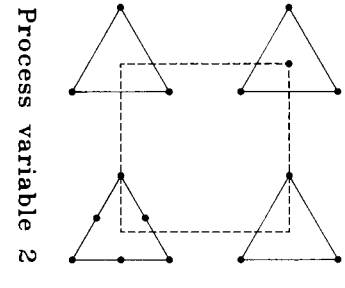


Fig. 2. Fractionated design for three components and two process variables. The four mixture triangles are depicted at the corners of a 2^2 design.

nent i). Projection as such is generally not sufficient to get feasible settings in the mixture variables. There are constraints on the mixture settings, namely the fractions have a minimum amount of 0. It is also possible that there are more restrictive constraints on the settings of the mixture variables, which results in a constrained mixture design [2]. To conform to these constraints the centre of the design projected is placed on the centre of the feasible design space and the design points are contracted towards this centre.

The projection of a 2^3 design is shown in Fig. 3. The bottom point of this figure is the origin. The cube is the 2^3 design with levels 0 and 1. The dashed triangle in the cube is the mixture triangle, conforming to $\Sigma\ m_i = 1$. The projection results in the following design points. The top (1, 1, 1) and the bottom (0, 0, 0) are projected on the centre (1/3, 1/3, 1/3). Three points already conform to the mixture plane (1, 0, 0), (0, 1, 0) and (0, 0, 1). The final three points (1, 1, 0), (1, 0, 1) and (0, 1, 1) are projected to (2/3, 2/3, −1/3), (2/3, −1/3, 2/3) and (−1/3, 2/3, 2/3), respectively.

The projected points have to conform to the constraints of the mixture plane. The centre of the design projected is the centre of the mixture design space, so it does not have to be moved. Contraction of all the design points to the centre (1/3, 1/3, 1/3) yields (2/3, 1/6, 1/6), (1/6, 2/3, 1/6) and (1/6, 1/6, 2/3) from (1, 0, 0), (0, 1, 0) and (0, 0, 1), respectively. The three points (2/3, 2/3, −1/3), (2/3, −1/3, 2/3) and (−1/3, 2/3, 2/3) result in (1/2, 1/2, 0), (1/2, 0, 1/2) and (0, 1/2, 1/2).

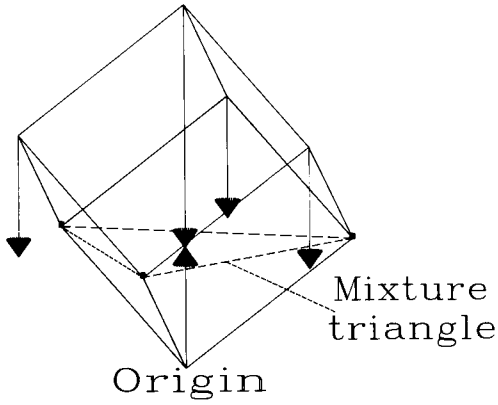


Fig. 3. Projection of a 2^3 design on a three-component mixture plane. The 2^3 design is used at levels 0 and 1. The triangle in the cube is the mixture triangle.

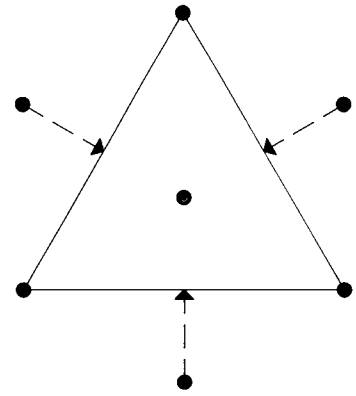


Fig. 4. Contraction of points which are outside the mixture triangle after projection. The points on the corners of the triangle are not contracted.

The disadvantage of this design is that the mixture design space is only partly covered. Therefore we propose a flexible contraction. This means that the points are contracted until they are inside the feasible region as is shown in Fig. 4. The mixture triangle shown here is also shown in Fig. 3 (dashed triangle). In Fig. 4 the projected design points are shown as dots. The three points outside the mixture triangle are contracted onto the perimeters of the mixture triangle. This results in the following design points: (1/3, 1/3, 1/3), (1, 0, 0), (0, 1, 0), (0, 0, 1), (1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2) and again (1/3, 1/3, 1/3). In our example, with two process variables and three mixture variables this results in the design shown in Fig. 5.

DESIGN EVALUATION

General

There are several approaches to the examination of the quality of experimental designs [6]. The most simple is to use optimal design criteria [7]. Then it is examined how correct a design is for estimation of the model parameters or predic-

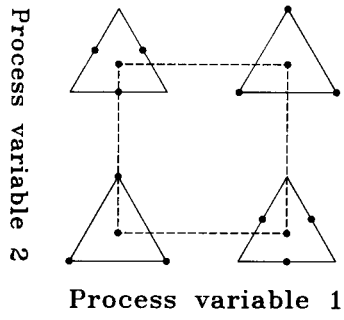


Fig. 5. Projection design for three components and two process variables. The four mixture triangles are depicted at the corners of a 2^2 design.

tions at predefined settings. Besides optimal design criteria it is also possible to use data for the evaluation of the quality of designs. Then the quality of results made on basis of the different designs is used. Here the prediction error of a test set as well as the model selection capabilities are of importance.

Prior knowledge

The optimal design criteria as well as other criteria which are suitable for the evaluation of designs can best be used if the prior knowledge is fully employed. We will use our designs in the optimization of directly compressed tablets; therefore we will use that prior knowledge.

The formulation of a tablet with direct compression is part of pharmaceutical technology. Tablets have several conflicting properties. They have to be strong to endure transportation and have to last during storage. Another property is that after consumption, a tablet should disintegrate and the drug itself should dissolve as fast as possible. To achieve these goals a tablet is composed from a drug to which various compounds are added, such as filler/binders, glidant, lubricant and disintegrant. The compounds are mixed and the tablet is constructed by compressing the powder resulting from the mixing process. This is a purely physical process. The properties of the tablet depend on the composition of the tablet and on the process variables such as mixing time and compression force. In this paper we use three filler/binders (α-lactose monohydrate, anhydrous β-lactose and primotab) as composition variables and compression force and mixing time as process variables. There are two responses that are measured, namely crushing strength and dissolution rate. More information about the specific experimental conditions can be found in the experimental section.

On basis of prior knowledge [8] we assume that the effect of both our process variables on the responses is linear in the range used. The mixture variables can have second or third-order blending effects. It is assumed that the blending properties of the compounds are independent of the level of the process variables, but this had not been tested experimentally.

Models

Prior knowledge results in a set of eight suitable models for three mixture variables and two process variables. These models all have different blending and interaction effects. The models are shown in Table 1. We present the models in table form because it allows a fast comparison of the models. In each model-table the top row contains a model for the process variables, while the left column contains a mixture model. The model terms indicated by the marks consist of the term of the top row multiplied by the term in the left column. We tried to construct hierarchical models, which means that higher-order interactions and blending effects are only used when the accompanying lower-order effects are present. The absence of marks in the row with "1" in the mixture model column results from the mixture restriction. Because of this restriction a process variable effect (or the intercept) cannot be estimated if the interactions between mixture effect and that process variable effect are calculated.

Design evaluation with theoretical criteria

Which criterion is most suitable for design evaluation depends on the purpose of the design. Although our purpose seems very clear (to find the settings with the optimum response), it does not result in a design criterion. This is because there is no simple method to calculate this optimum. The method to be used consists of several parts. We first examine the data, try several models for suitability and then decide which to use.

TABLE 1

The eight models used
(A model term consists of the mixture model term multiplied by the process variable model term. Each model consists of the terms indicated by the marks. Models 2 and 6 are mirror images in the process variables.)

1	1	mt	cl	mt*cl
m_1	x	x	x	
m_2	x	x	x	
m_3	x	x	x	
m_{12}				
m_{13}				
m_{23}				
m_{123}				

2m	1	mt	cl	mt*cl
m_1	x	x		
m_2	x	x		
m_3	x	x		
m_{12}	x			
m_{13}	x			
m_{23}	x			
m_{123}				

2c	1	mt	cl	mt*cl
m_1	x		x	
m_2	x		x	
m_3	x		x	
m_{12}	x			
m_{13}	x			
m_{23}	x			
m_{123}				

3	1	mt	cl	mt*cl
m_1	x	x	x	
m_2	x	x	x	
m_3	x	x	x	
m_{12}				
m_{13}				
m_{23}				
m_{123}				

4	1	mt	cl	mt*cl
m_1	x	x	x	
m_2	x	x	x	
m_3	x	x	x	
m_{12}	x			
m_{13}	x			
m_{23}	x			
m_{123}				

5	1	mt	cl	mt*cl
m_1	x	x	x	
m_2	x	x	x	
m_3	x	x	x	
m_{12}	x			
m_{13}	x			
m_{23}	x			
m_{123}	x			

6m	1	mt	cl	mt*cl
m_1	x	x		
m_2	x	x		
m_3	x	x		
m_{12}	x	x		
m_{13}	x	x		
m_{23}	x	x		
m_{123}				

6c	1	mt	cl	mt*cl
m_1	x		x	
m_2	x		x	
m_3	x		x	
m_{12}	x		x	
m_{13}	x		x	
m_{23}	x		x	
m_{123}				

Maybe we examine the residuals and check for outliers. Then, when the model is accepted we calculate the response surface and try to find the optimum. If there are several responses we have to use all response surfaces to get an optimum alternative.

This shows that there are several limitations to the optimal design criteria. If the model is not known beforehand then it is impossible to decide which design would result in best estimations of the parameters or in best predictions of the response. Therefore in the assessment of the designs we chose to use six models and present results for all models. Only six from the previous eight models were used since two pairs of the models are mirror images in the process variables. These pairs of models have the same results in the evaluation since the designs are symmetrical in the process variables. Several design properties are important for us, such as precision of the parameters, precision of the predictions (depending on the settings of the independent variables). We will use the average variance of the parameters and the average variance of predictions to compare the designs. These values are calculated with a measurement variance of $\sigma^2 = 1$.

To explain the calculations necessary for the design evaluation criteria some introductory formulas are given. Then the formulas for the variance of the parameters will be presented. The following notation is used.

- $\mathbf{X}$ Design matrix with settings and interactions of the model
- $\mathbf{x}$ Vector with setting and interactions for one design point
- y One response
- $\mathbf{y}$ Vector with all responses
- $\boldsymbol{\beta}$ Vector with parameters
- ϵ Error in response
- tr trace of a matrix
- $'$ transpose operator

The model is:

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon$$

and the model parameters are calculated by:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

The two criteria for the examination of the designs are:

$$A = \mathrm{tr}(\mathbf{X}'\mathbf{X})^{-1}$$

(error in parameter estimation)

$$V = \underset{\mathbf{x}}{\text{mean}}\ \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$$

(prediction error averaged over $\mathbf{x}$)

Good designs are those that minimize the value of A and V. To get the mean variance of the parameters the A values are divided by the number of parameters [7].

The V criterion is used to examine the predic-

tion error. The prediction error is a measure of the precision with which prediction at a certain point can be made. The mean prediction error is the error which is expected when predicting a response at unknown factor settings in the design space. The V criterion is calculated over a grid of design points. The grid used has 25 different settings for the process variables and 31 different compositions. In the mixture space the grid had a point distance 0.2, this resulted in 21 design points. Furthermore, "interesting points" in the mixture space were added. These were binary mixtures with composition (1/2, 1/2, 0), binary mixtures with composition (2/3, 1/3, 0) and a ternary mixture with composition (1/3, 1/3, 1/3). This resulted in another 10 compositions. In the process variable space five levels (−1, −0.5, 0, 0.5, 1) were used for both process variables. Thus 25 factor settings were used and combined with the 31 mixture settings. This resulted in $25 \cdot 31 = 775$ evaluation points.

Design comparison on practical data

Several properties of the data are of interest for the comparison of the designs generated. These properties are related to the two purposes of the design, the selection of a model and the calculation of the model parameters. The questions are more specifically:

- How trustworthy is the model selection?
- What is the expected prediction error from the design–model combinations?
- What is the prediction error on the test set?

The model selection cannot be evaluated with knowledge of the "true" model (the model describing the state of nature) since that model does not exist [5]. Therefore the prediction error on a test set is used to evaluate the model selection. This prediction error is calculated as mean squared error of prediction test-set (MSEPT). The MSEPT of a model can of course be compared with the measurement error. However, we do not have genuine replicates [6], but replicated tablets from the same batch. Thus we have an idea of the measurement error, but, since not all sources of error are included, the true measurement error is larger. Some model selection criteria, such as prediction criterion (PC) [9] and generalized cross-validation (GCV) [10] are estimates of the MSEPT. We will use that property in our examination. We will first mention the problems associated with the usage of experimental data for this purpose and then explain our model selection criteria.

When using data to evaluate the quality of designs there are pitfalls which one must avoid, or at least be aware of. The main danger is getting conclusions which are too much dependent on these data. So the conclusions may not be generalized to other situations. This can be caused by two things. The first is that the responses used are not representative of all possible responses. The second is the problem of outliers: completely wrong responses at design points. Therefore we refrain from making general statements about the validity of the results.

MODEL SELECTION

Our model selection procedure is fairly simple. We use eight different models and calculate two model selection criteria for all models. The model which is selected does best on the basis of both model selection criteria. If one model has the lowest value for both criteria then that model is selected. If such is not the case then an ambiguous situation appears and no clear decision is made. We refrain from using a more complex and flexible method (such as stepwise model selection [11,12]) since it is more difficult to implement [2] and it increases the effects of data mining [12,13].

There is a wealth of criteria available for the examination of the fit of a model to a set of data [11,14]. Most of them do not only result in a model preferred (lowest or highest criterion value), but can also be interpreted as estimators of the MSEP. Our interest is directed to the prediction error of models and thus we prefer model selection methods which measure this property. The prediction error of a model is usually calculated in terms of MSEP. In our case MSEP is the mean prediction error averaged over the design space.

As a first criterion we chose the PC of Amemiya [9], which uses the residual variance of

the model for the assessment of the model quality. The PC estimates the MSEP of a model with the assumption that the points (x settings) at which the predictions are made are random and have the same distribution as the design points. It is calculated by:

$$PC = \left(1 + \frac{p}{n}\right) s^2$$

with p the number of parameters, n the number of experiments and s^2 the residual variance.

We use GCV [10] as a second model evaluation criterion. It estimates the MSEP on the design points. Cross-validation methods are based upon the data-splitting principle. The estimated prediction error obtained from cross-validation or generalized cross-validation is expressed as PRESS. To calculate the PRESS all design points are successively removed from the design while their responses are predicted on basis of the remainder of the design points. The resulting prediction errors are squared and added together.

There are several disadvantages to cross-validation when a very small design (small compared to the model) is used. The first is that when a design point is removed it is sometimes impossible to calculate the model parameters and thus the predicted y value anymore. A problem of the PRESS is that design points with a high leverage may have disproportionate high residuals when they are left out, which results in too high a PRESS value. Examples of these points are the design points on the edge of the design space, where the model is extrapolated for the calculation of a PRESS prediction. A third disadvantage of a cross-validation method is that it is not known how many degrees of freedom are associated with the estimator.

To overcome the leverage problem Golub et al. [10] have constructed the GCV, which is PRESS-corrected for the leverage of the design points. This means that the residuals of points with a low leverage are multiplied by a larger quantity, whereas residuals of points with a higher leverage get a lower quantity. Division of the GCV value by the number of data points results in a mean GCV (MGCV). The GCV does not solve the problem of the unknown number of degrees of freedom. The difference between MPRESS and MGCV is most profound if there are large differences between the leverages of the design points; if all design points have equal leverages then MPRESS and MGCV are equal. GCV is calculated via:

$$\mathrm{GCV} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\setminus i}\right)^2 \frac{1 - h_i}{1 - p/n}$$

with $\hat{y}_{\setminus i}$ the prediction on design point i without using this observation, h_i the leverage of design point i, p the number of parameters in the model and n the total number of observations. To calculate the MGCV the GCV is divided by n, the number of observations.

MGCV and PC are comparable measures. They can also be compared with MSEPT, the MSEP estimated on a test-set. To calculate the MSEPT the model is calculated on the current experimental design (training set) and another set of data (test-set) is used to calculate the mean squared prediction error. We use the same test-set for all our designs. This test-set consists of all points with composition 1/3 of a component, 2/3 of another component and nothing of the remaining third component. These six compositions settings are repeated on all levels of the process variables, thus resulting in a test-set of $6 \cdot 2 \cdot 2 = 24$ design points. The formula for the MSEPT is:

$$\mathrm{MSEPT} = 1/n_t \sum_{i=1}^{n_t} \left(y_{t,i} - \hat{y}_{t,i}\right)^2$$

with n_t the number of observations in the test-set, $y_{t,i}$, observation i from the test-set and $\hat{y}_{t,i}$ the prediction at observation i of the test-set.

We chose our specific model selection method to avoid the problem of data mining [12,13]. Since we test only a few models we avoid chance correlations which make the model validation criteria (estimated prediction error) too optimistic.

EXPERIMENTAL

Composition of the tablets

Three filler/binders were used as components of which the fractions had to be optimized. These

were α-lactose monohydrate (100 mesh, DMV, Veghel, Netherlands), anhydrous β-lactose (marketed as Pharmatose® 21, DMV) and modified rice starch (developed by Erawan Pharmaceutical Research and Laboratory, Bangkok, and marketed as Primatab® by Avebe, Foxhol, Netherlands). The drug, disintegrant, glidant and lubricant were used in a constant concentration of 4%, 4%, 0.2% and 1%, respectively. The tablets were lubricated with magnesium stearate (Pharm. Eur. grade, OPG, Utrecht), which was sieved through a 210-μm sieve prior to use. The disintegrant used was Primojel (Avebe). Aerosil 200® (Degussa, Frankfurt) was sieved through a 210-μm sieve and was added as a glidant. The drug used was oxazepam (Pharm. Eur. grade, Pharmachemie, Haarlem).

Procedure

All ingredients except magnesium stearate were mixed for 15 min in a Turbula mixer (Model 2p, W.A. Backhoven, Basle) at a rotation speed of 90 rpm. Magnesium stearate was added and the mixing was continued for 2 or 10 min (first process variable, low and high level). Tablets (250 mg) were prepared on a single-punch tabletting machine (HOKO, Rijswijk, Netherlands), using 9-mm flat punches. Tablets were produced at a compression load of 157 MPa or 314 MPa (second process variable). Unfortunately the compression load level could not be set with the preferred precision, however, it was possible to measure the compression load of the tablets. In our calculations we used the average compression load. The tabletting was performed in a room with constant temperature (20 ± 1°C) and relative humidity (45 ± 5%).

Tablet property

Crushing strength. From each batch the crushing strength of 10 tablets was measured using a Schleuniger instrument (Model 4M, Dr. K. Schleuniger, Zürich). The average crushing strength of these 10 tablets was used as response value in the design point.

Dissolution rate. The dissolution rate measurements of oxazepam from the tablets were performed using the USP XXI paddle method, at 50 rpm. As dissolution medium 1000 ml of deaerated water (37 ± 1°C) was used. The concentration of oxazepam was measured spectrophotometrically at 231 nm (Ultrospec 4052 TDS, LkB, Zoetermeer, Netherlands). The measurement was performed automatically, monitored by an Apple II computer. Samples were taken after 10 min. From each batch the dissolution rate of six tablets was measured. The average dissolution rate of these six tablets was used as response value in the design point.

RESULTS

Theoretical evaluation

The theoretical design evaluation is not performed for all the models. This is because of the symmetry between the process variables. This symmetry causes the models 2c and 2m and models 6c and 6m to have the same theoretical values. In the evaluation on experimental data all models are used since mixing time and compression force have different effects on the responses.

The first result of the theoretical examination is that with the fractionated design model 6 cannot be calculated. This is a disadvantage of this design. The first criterion for the theoretical evaluation was the A criterion (Fig. 6), the average variance of the parameters of the model. It is

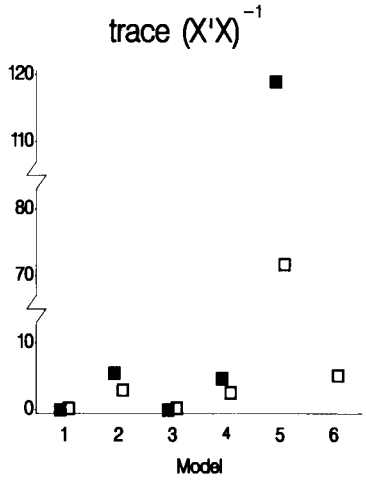


Fig. 6. Values of the A criterion vs. the models. (■) Fractionated design; (□) projected design.

clear that there are quite some differences between the different models.

According to the A criterion the projection design is better with models 2, 4 and 5. Models 1 and 3 can better be estimated with the fractionated design. The difference between the A criterion values of the designs is large enough to be of significance.

For the examination of the expected prediction variance the V criterion (Fig. 7) is used. The V criterion for the flexible projected design does not change much between the different models. The fractionated design gives much larger differences for the different models than the projected design. With models 1 and 3 the fractionated design is much better than the projected design. For models 2, 4 and 5 the fractionated design performs worst. The good results of the fractionated design with models 1 and 3 can be explained from the set-up of the designs. This design has design points at all pure components and all process variable settings. This means that the linear blending effects and the interaction of the linear blending effects with the process variables can be calculated relatively precise. In the projection design only at half of the process variable levels measurements at the pure components are scheduled. At the other half of the process variable levels there are measurements at binary mixture compositions. With those binary mixtures

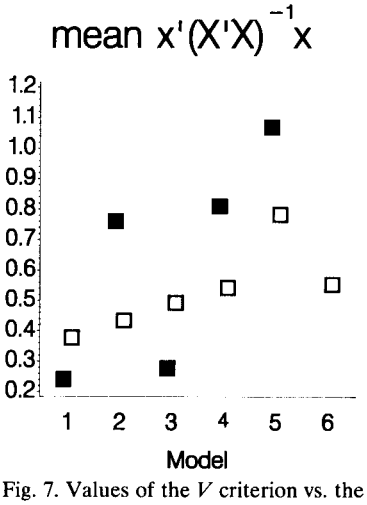


Fig. 7. Values of the V criterion vs. the models. (■) Fractionated design; (□) projected design.

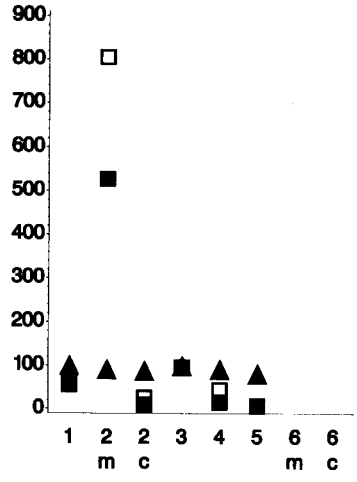


Fig. 8. Regression results of the fractionated design with crushing strength data. (■) PC; (□) MGCV; (▲) MSEPT.

the non-linear blending effects can be calculated more precisely. This is at the expense of the precision of calculation of the linear blending effects and the interaction between the linear blending effects with the process variables.

Results with measured data

When experimental data is used, it appears that not all models can be cross-validated. With the fractionated design models 6m and 6c cannot be calculated, and for models 3 and 5 only the PC is available to validate the model with. This is a disadvantage for this design.

For the fractionated design and using crushing strength (Fig. 8) it seems that all models have about the same MSEPT values. The PC value of model 5 is the lowest; this model has also the lowest MSEPT value. If a model which can be cross-validated is preferred, then model 2c is the best. It has the next lowest PC value and the lowest MGCV value. If this model is used then the second best MSEPT is found.

With dissolution rate and the fractionated design (Fig. 9) model 5 again has the lowest PC value. In fact its PC value is so low (3.8) that we do not trust it. If that model is rejected because it cannot be cross-validated, then model 4 has the next lowest PC value and the lowest MGCV value. Model 4 also has the lowest MSEPT value. Model 5, which has the lowest PC value, has an

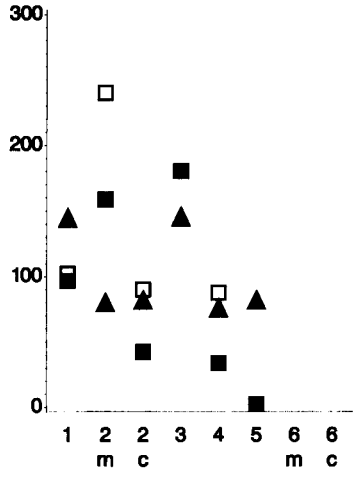


Fig. 9. Regression results of the fractionated design with dissolution rate data. (■) PC; (□) MGCV; (▲) MSEPT.

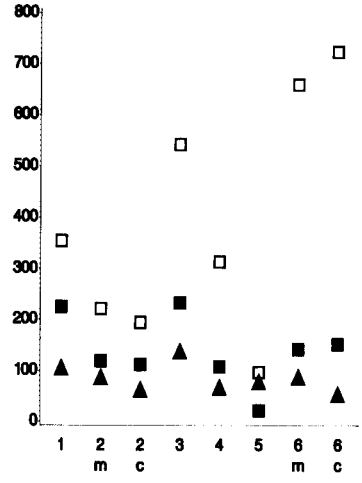


Fig. 11. Regression results of the projected design with dissolution rate data. (■) PC; (□) MGCV; (▲) MSEPT.

MSEPT (84 kg^2) which is slightly larger than the MSEPT of model 4 (78 kg^2).

Therefore, with the fractionated design the acceptable models (in terms of the prediction error) can be selected. The MSEPT of crushing strength of model 5 is 85 kg^2 which is relatively large compared to the measurement error, which is about 43 kg^2. The dissolution rate has a lowest MSEPT of 77%^{2} while the measurement error is 27%^{2}.

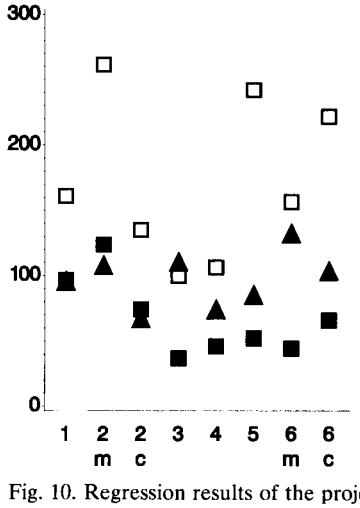


Fig. 10. Regression results of the projected design with crushing strength data. (■) PC; (□) MGCV; (▲) MSEPT.

All models can be calculated and cross-validated with the projected design. With crushing strength (Fig. 10) model 3 has the lowest PC and MGCV values (38 and 101 kg^2, respectively). This model has too high a MSEPT value (112 kg^2) to be acceptable. The best model (if the MSEPT is considered) is model 2c. This model has an MSEPT value of 69 kg^2 and would not be selected on basis of the data.

With dissolution rate and the projection design (Fig. 11) model 5 has the lowest PC value (25%^{2}) This model also has the lowest MGCV value (100%^{2}) but is does not have the lowest MSEPT value (82%^{2}). The model with the lowest MSEPT is model 6c (67%^{2}), which is not different enough to be of interest. This model has the highest MGCV value and it would never have been selected on basis of the data.

Therefore, with the projection design the models selected are not always acceptable. The prediction errors of the best performing models are lower than the prediction errors of the best performing models of the fractionated design.

Conclusion

The designs can result in acceptable predictions. The projection design has somewhat lower

prediction errors on the test-set than the fractionated design. This can be explained by the fact that the projection design had better V criterion values for the models based on the experimental data than the fractionated design.

The model selection with the projection designs is unsatisfactory. With the fractionated design the model selection is somewhat better. If there is knowledge about the model to be used then the best performing design on that model can be used. If, in that case, there are more responses, preferably the projection design can be used. Otherwise larger designs should be preferred.

The designs examined here are intended for our own situation, where the assumptions we made are true. Under other circumstances our conclusions are not valid and other designs and design construction techniques should be used. If there are restrictions on the mixture variables, resulting in an irregular shaped design space, neither of the designs is feasible and then an optimal design [7] (not examined here) should be used.

The authors wish to express their gratitude towards the following companies: DMV, FMC, Erawan and Pharmachemie for the donation of materials (α- and β-lactose, Primatab, Primojel and oxazepam, respectively). This research is supported by the Netherlands Technology Foundation (STW).

REFERENCES

1 C.E. Bos, G.K. Bolhuis, C.F. Lerk, J.H. De Boer, C.A.A. Duineveld, A.K. Smilde and D.A. Doornbos, Drug Development and Industrial Pharmacy, Vol. 17, No. 18, 1991, pp. 2477–2496.
2 J.A. Cornell, Experiments with Mixtures, Wiley, New York, 2nd end., 1990.
3 J.A. Cornell and J.W. Gorman, J. Qual. Technol., 16 (1984) 20.
4 V. Czitrom, Commun. Stat. Theory Methods, 17 (1988) 105.
5 I. Hau, PhD Thesis, University of Wisconsin, Madison, WI, 1990.
6 G.E.P. Box and N.R. Draper, Empirical Model-Building and Response Surfaces, Wiley, New York, 1987.
7 V.V. Fedorov, Theory of Optimal Experiments, Academic Press, New York, 1972.
8 C.E. Bos, personal communication.
9 T. Amemiya, Advanced Econometrics Basil, Blackwell, London, 1985.
10 G.H. Golub, M. Heath, and G. Wahba, Technometrics, 21 (1979) 215.
11 R.R. Hocking, Biometrics, 32 (1976) 1.
12 A.J. Miller, Subset Selection in Regression, Chapman and Hall, London, 1990.
13 M.C. Lovell, The Review of Economics and Statistics, 65 (1983) 1.
14 G.G. Judge, W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.-C. Lee, The Theory and Practice of Econometrics, Wiley, New York, 2nd end., 1985.

Analytica Chimica Acta, 277 (1993) 467–471
Elsevier Science Publishers B.V., Amsterdam

Prediction of mixture composition by chromatographic characterization, multivariate classification and partial least-squares regression, a comparison of methods

Peter Kaufmann

Karlshamns LipidTeknik AB, P.O. Box 6686, S-113 84 Stockholm (Sweden)

(Received 3rd September 1992)

Abstract

Elucidation of the mixing proportions of mixtures of natural triacylglycerol oils is of paramount importance in several areas, most notably in quality control in the pharmaceutical, cosmetic and food industries. The detection of adulteration of edible oils, i.e., where high-priced commodity oils are mixed with lower-priced substitutes, is also of primary importance. Three goals were set for this study; to construct a model that detects adulteration, that identifies the components in the mixture and also predicts the levels of the components present in the mixture. This was done by utilizing the fatty acid profiles and the triacylglycerol molecular species profile, analyzed by chromatographic methods, principal components analysis and partial least-squares regression. All three goals of this study were fulfilled.

Keywords: Chromatography; Principal component analysis; Mixture composition prediction; Quality control; Triacylglycerol oils

The concept of identity is of utmost importance in quality control and quality assurance in, among others, the pharmaceutical, cosmetic and food industries. Incoming batches of raw materials must be tested for compliance to specifications prior to processing. Finished products must also be tested for compliance to product specification, which also can be seen as identification. This concept can be extended to include cases where admixture of higher priced commodities with lower priced substitutes occurs, i.e., adulteration. This last aspect holds vital implications for the consumer of edible oils, as the "Spanish toxic syndrome" has shown [1]. In all these instances it is of interest not only to identify the components but also to quantify at which levels they are present.

Correspondence to: P. Kaufmann, Karlshamns LipidTeknik AB, P.O. Box 6686, S-113 84 Stockholm (Sweden).

The overall dominating lipid class (> 90%) present in most natural oils are the triacylglycerols (TAGs), which are composed of three acyl chains (fatty acids) esterified to a glycerol backbone. Natural oils can be characterized by their fatty acid and triacylglycerol profiles, which can be measured by gas chromatography (GC) [2] and liquid chromatography (LC) [3], respectively. These profiles exhibit differing degrees of complexity depending on the source and thus furnish a more or less unique means of identification.

The fatty acid and/or the TAG profiles have been used in several applications to identify or predict the levels of components in mixtures. Selected fatty acids, measured by GC, in conjunction with multiple regression have been used to detect cow milk in buffalo milk down to the 5% level [4]. Different TAG molecular species, measured by GC, and multiple linear regression were used to detect the presence of seven different

vegetable oils and two different animal fats in cow milk [5], with a later improvement and refinement to a lower level of detection (4–7%) and extension of the number of fats and oils to twenty [6]. The total TAG profile, measured by LC, was used in conjunction with principal components analysis to detect adulteration of four different vegetable oils in olive oil to between the 10 and 20% level [7]. Butteroil in margarine was quantified utilizing selected fatty acids, measured by GC, and multiple regression [8]. The fatty acid, sterol and tocopherol profiles were used together with multiple regression and statistical tests to develop a system that selects the oils to be included in the model, thus enabling accurate predictions [9]. The total fatty acid and the TAG profiles, measured by GC and LC respectively, were used in conjunction with principal components analysis and SIMCA classification to identify and detect adulteration in vegetable oils [10].

The goals of this study were to construct a method, utilizing chromatographic characterization, that provides answers to the following questions; (a) is the oil adulterated, (b) with which oil or oils is the sample adulterated with, and (c) at what levels are the different components present in the mixture.

EXPERIMENTAL

All solvents and reagents were of analytical grade or better and were obtained from Merck (Darmstadt).

Gas chromatography

TAGs from the oils were transesterified to fatty acid methyl esters (FAME) according to Ref. 2. These were subsequently analyzed by GC, also according to Ref. 2. In brief: a DB-Wax column, 30 m × 0.25 mm i.d. (J&W, Folsom, CA) was used, the oven was programmed from 130°C (hold 2 min) to 150°C at 50°C min^{-1}, then to 235°C at 3.3°C min^{-1} and subsequently held at the final temperature for 22 min.

Liquid chromatography

25 mg of the oil was dissolved in 100 ml of ethanol–isooctane (3:1, v/v). The TAG molecular species profile was analyzed according to Ref. 3. In brief: a Lichrocart 250 mm × 4 mm i.d. Superspher 100 ODS 4 μm column from Merck was used, with a gradient of 100% acetonitrile to 100% acetonitrile–ethanol–isooctane (4 : 4 : 2, w/w/w) in 60 min with a negative exponential profile. Detection was done by a Cunow DDL 11 light-scattering detector (Cergy St. Christophe, France) operating at a pressure of 30 p.s.i. and a temperature of 80°C.

TABLE 1

Characteristics of the oils

Oil	Number of fatty acids	Number of TAG species	n [a]	$S.D._{nat.}$	$S.D._{sim.\ set}$
Linseed	10	22	1	–	0.236
Olive	9	11	3	0.328	0.395
Palmkernel	9	21	1	–	0.247
Peanut	11	15	1	–	0.199
Safflower	10	17	3	0.404	0.409
Sesame	10	13	1	–	0.275
Soybean	6	16	4	0.221	0.236
Sunflower	10	12	2	1.895	1.482
Sunflower HO (high oleic)	9	12	2	0.263	0.303

[a] Number of authentic samples.

Data

In order to build principal component models according to Refs. 11 and 12 for each of the oil types (Table 1) a semi-simulated data-set was generated. For oil types with two or more authentic samples (Table 1) this was done by using the natural standard deviation (S.D.) for each of the peaks present, multiplying this by a random number and randomly adding or subtracting these quantities from the average chromatogram of each oil type, thus generating 14 new samples. This procedure, New peak = Authentic peak ± (rand · S.D.), generated populations with standard deviations close to the natural standard deviations (Table 1). For oil types with only one authentic sample a new sample was generated by randomly adding or substracting to peaks in the authentic sample and thereafter using the above described procedure, thereby producing populations with comparable standard deviations.

Nine different calibration mixtures (50:50, 60:40, 70:30, 80:20, 90:10, 95:5, 97:3, 98:2 and 99:1) of four different oils, peanut in olive, safflower in olive, sesame in olive and sunflower HO in olive, were used for the calibration set. These were created by summing chromatograms of the authentic oils multiplied by the appropriate proportionality factor. A test-set of eight mixtures (Table 2) from the calibration set was withheld from the model-building processes in order to validate the model by calculating the prediction error.

Classification

Individual principal component models for each of the nine oil types (Table 1) were developed using the SIMCA 3B program (Sepanova AB, Stockholm). The theory and practice of these techniques have been described in detail in the chemometric literature [11,12]. Each of the calibration mixtures described above and also each of the authentic oils were fitted to each of the nine class models. The degree of fit to each model, described by the residual standard deviation (R.S.D.) of each oil sample, for both the calibration set and the authentic oil samples was used in building the prediction model.

TABLE 2

Prediction error [a], test-set, safflower, peanut, sesame or sunflower HO in olive

Mixture	Class R.S.D. (vol.%)		Chromatographic var. (vol.%)	
	Olive	Other [b]	Olive	Other [b]
1. Olive 97:3 safflower	4.8	16.2	0.3	16.2
2. Olive 99:1 safflower	4.2	67.0	0.3	52.0
3. Olive 60:40 peanut	11.3	50.5	0.1	0.5
4. Olive 80:20 peanut	5.1	76.6	0.1	1.0
5. Olive 70:30 sesame	17.0	55.7	0.4	0.7
6. Olive 95:5 sesame	1.26	54.8	0.2	8.0
7. Olive 80:20 sunflower HO	3.5	15.0	0.3	1.0
8. Olive 95:5 sunflower HO	2.0	54.0	0.3	8.6

[a] Relative prediction error. [b] Minor component.

Prediction

Prediction of mixture composition was done by partial least-squares regression (PLS), a technique widely used in the chemometric literature

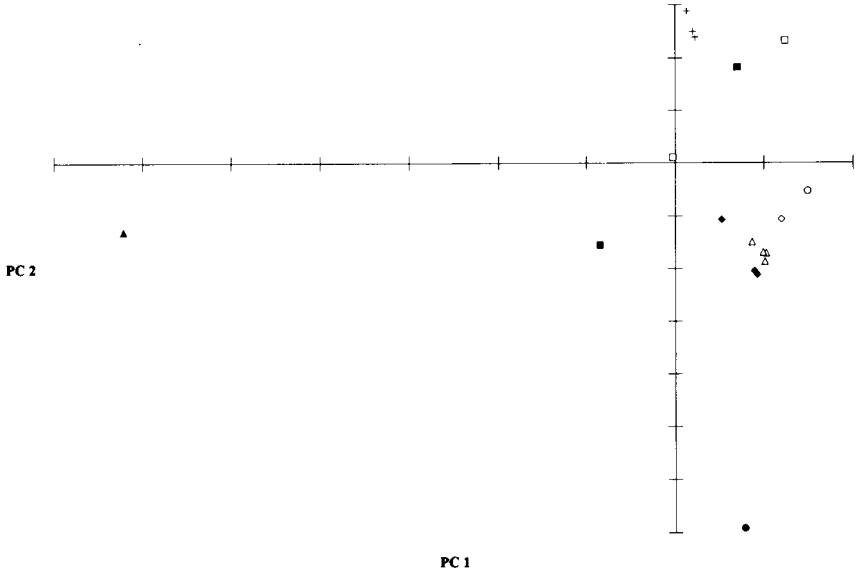


Fig. 1. Plot of the first two principal components of all chromatographic variables for the authentic oils. □ = Sunflower HO; ■ = sunflower; △ = soybean; ◇ = sesame; ◆ = safflower; ○ = peanut; ▲ = palmkernel; + = olive; ● = linseed.

for multivariate calibration, the theory and practice of which has been described in detail [12–14]. Two different calibrations were undertaken, one where the R.S.D. values for the calibration set and for the authentic oils were taken as the independent variables (X-block) and the matrix of mixing proportions as the dependent variables (Y-block). In the second calibration the chromatographic data for the authentic samples and also for the calibration set were used as the independent variables and the matrix of mixing proportions as the dependent variables.

TABLE 3

Multivariate classification

Oil	Number of significant components	Variance explained by model (%)
Linseed	1	73.0
Olive	3	69.0
Palmkernel	2	75.5
Peanut	1	76.0
Safflower	2	67.0
Sesame	1	69.0
Soybean	3	72.0
Sunflower	3	71.0
Sunflower HO	2	71.0

RESULTS AND DISCUSSION

A plot of the first two principal components shows the spread of the different authentic types of oils in the chromatographic measurement space. As can be seen in this plot, palmkernel oil deviates greatly from the other oils in both PC1 and PC2, linseed oil and olive oil also differ greatly, although mainly along the PC2 axis (Fig. 1). The two different types of sunflower oil are intermingled (Fig. 1). The remaining oils are grouped together in the lower right-hand quadrant. These separations are perhaps reminiscent of taxonomical similarities.

Multivariate class models were developed for each of the oils (Table 1), the number of significant principal components were determined by crossvalidation [15], the number of components and the variance explained by each class model are shown in Table 3. Thereafter each of the samples in the calibration set and the authentic oils were fitted to the class models and their R.S.D. calculated. The R.S.D. is a measure of the

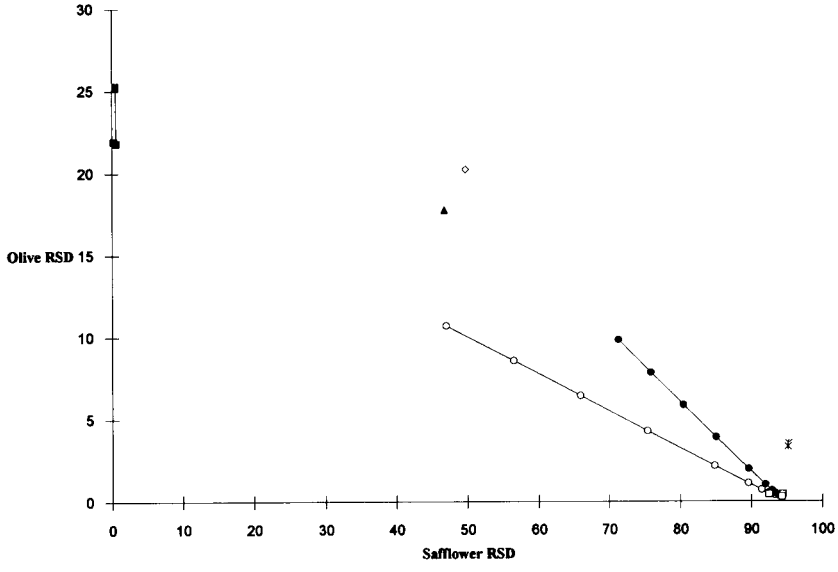


Fig. 2. Plot of the class R.S.D. for safflower oil and olive oil, showing the linear progression of mixtures from one pure class to the other pure class, according to mixing proportions. ○ = Olive/Safflower; ● = olive/peanut; □ = olive; ◇ = peanut; ■ = safflower; ▲ = sesame; * = sunflower HO.

fit to the class model for each object tested, essentially the distance from the class model in the multidimensional measurement space [12]. A plot of R.S.D. values for the classes olive oil and safflower oil shows an encouragingly clear and linear progression from the one oil class to the other, according to mixing proportions (Fig. 2). The R.S.D. values for the training-set were used as *X*-block variables and the corresponding matrix of mixing proportions as the *Y*-block in the subsequent PLS regression [12]. A nine-component PLS model was developed explaining 99% of the variance in the *X*-Block and 79% of the variance in the *Y*-block. The crossvalidation [15] criteria for the selection of components for this model were not applied as strictly as for the classification models. Several mistaken identifications of the oils present in the different mixtures in the test-set occurred. As can be seen in Table 2 the relative prediction errors for this PLS model exhibit more or less acceptable predictions for the major component in the mixture while the minor components are grossly mispredicted (Table 2).

The second prediction trial used a more classical approach, where the chromatographic variables for the calibration set and the authentic oils were used as the *X*-block and the corresponding matrix of mixture proportions were used as the *Y*-block. Nine significant PLS components according to crossvalidation [15] were used in the model, explaining 93% of the variance in the *X*-block and 96% of the variance in the *Y*-block. Identifications of the oils present in the mixtures for the test-set were all correct for this model. The prediction errors for this model are given in Table 2, as can be seen the results for the major components are excellent, all well under 1%. The results for the minor components are also acceptable, with one exception (Table 2). However, if all predictions are rounded off to whole percents all levels are correctly predicted with the exception of safflower in olive, 99 : 1.

Conclusions

The results of this study show that, despite the promising appearance of Fig. 1, using class R.S.D. values to predict mixture composition does not seem to be feasible. However, using a more "classical" approach in building the prediction model gave excellent results for all three goals set for this study. Adulteration was detected down to at least the 1.5% level and the components in the mixtures were correctly identified. Prediction of the levels of the components in the mixtures were also correct down to the 1.5% level.

Thanks to Mr. M. Bergqvist, Dr. B.G. Herslöf and Dr. N.U. Olsson for use of chromatographic data.

REFERENCES

1 S.A. Jimeno, Trends Anal. Chem., 1 (1982) 4.
2 U. Olsson, P. Kaufmann and B.G. Herslöf, J. Chromatogr., 505 (1990) 385.
3 B.G. Herslöf, in S. Watanabe (Ed.), Proceedings ISF-JOCS World Congress 1988, Japan Oil Chemists Society, Tokyo, 1989, p. 892.
4 R.S. Farag, M.M. Hewedi, S.H. Abo-Raya and H.H. Khalifa, Fette Seifen Anstrichm., 88 (1986) 106.
5 D. Precht and K. Heine, Milchwissenschaft, 41 (1986) 406.
6 D. Precht, Z. Lehensm. Forsch., 194 (1992) 107.
7 M. Tsimidou, R. Macrae and I. Wilson, Food Chem., 25 (1987) 251.
8 G.L. Christen, J. Food Qual., 11 (1989) 453.
9 P.J. Van Niekirk and R.A. Hasty, Anal. Chim. Acta, 223 (1989) 237.
10 P. Kaufmann and B.G. Herslöf, Fat Sci. Technol., 93 (1991) 179.
11 S. Wold, Pattern Recogn., 8 (1976) 127.
12 S. Wold, C. Albano, W.J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson and M. Sjöström, in H. Martens and H. Russwurm (Eds.), Food Research and Data Analysis, Applied Science Publishers, London, 1983, p. 147.
13 H. Wold, in K.G. Jöreskog and H. Wold (Eds.), Systems Under Indirect Observation, North-Holland, Amsterdam, 1982, p. 307.
14 A. Lorber, L.E. Wangen and B.R. Kowalski, J. Chemometr., 1 (1987) 19.
15 S. Wold, Technometrics, 20 (1978) 397.

Analytica Chimica Acta, 277 (1993) 473–475
Elsevier Science Publishers B.V., Amsterdam

Bayesian slippage test for detection of outlying sub-samples

Ashok K. Singh

Department of Mathematics and Harry Reid Center for Environmental Studies, University of Nevada, Las Vegas, NV 89154 (USA)

Anita Singh

Lockheed Environmental Systems & Technologies Co., Environmental Sciences & Technologies Division, 980 Kelly Johnson Drive, Las Vegas, NV 89119 (USA)

(Received 3rd September 1992)

Abstract

A data set which can be divided into $k \geq 2$ distinct sub-samples is considered. The problem of testing of hypothesis of homogeneity of means or variances of the k sub-populations arises, for example, in estimation of background contamination in soil and groundwater in environmental investigations. We have considered the problem from a Bayesian approach. Cut-off points are computed for a Bayes test for testing the null hypothesis that all sub-samples have the same means versus the slippage alternatives that $t < k$ of these sub-samples have higher means. Some examples are given.

Keywords: Bayesian slippage test; Outlier test

An outlying observation or "outlier" has been defined as "one that appears to deviate markedly from other members of the sample in which it occurs". Clearly, our feelings about the data in this respect depends heavily on our choice of the underlying probability model. For example, an observation which appears to be extreme under the assumption of normality may not appear to be extreme under the assumption of log-normality. An outlier test is a statistical procedure for objectively determining if a sample can be considered to have come from one underlying population, or if it contains a few spurious observations. There are different reasons for the presence of outliers: (i) an error in a measurement, recording, or calculation, (ii) inexplicable random fluctuations, (iii) an error in sampling, etc. For a detailed discussion of categorizations of outliers, see Refs. 2–4.

A slightly different situation arises where a data set is composed of distinct sub-samples, and the goal is to statistically determine if some of the sub-samples have a mean (or variance) which is different from the remaining sub-samples. A special case of this problem is the problem of slippage testing. Mosteller [5] first considered the problem of testing homogeneity of k continuous populations against the alternative that one of these populations has a higher location parameter than the rest. Karlin and Truax [6] considered the problem of single slippage from the Bayesian approach, and derived a class of symmetric Bayes tests for a parametric case, and a class of locally optimum tests based on ranks in some nonparametric situations. Singh [7] extended the results of Karlin and Truax [6] to the case of multiple

Correspondence to: A.K. Singh, Department of Mathematics and Harry Reid Center for Environmental Studies, University of Nevada, Las Vegas, NV 89154 (USA).

slippages. In this paper, we numerically compute the cut-off points for the multiple slippage Bayes test of Singh [7] for the case of normal populations.

The multiple slippage testing problem arises, for example, in an environmental investigation dealing with soil and groundwater contamination, when the site is inhomogeneous with respect to soil type. In such a situation, the average background contamination may be different for the different soil types, and a test may be needed to statistically verify this inhomogeneity.

MATHEMATICAL FORMULATION OF THE PROBLEM

Let $\{x_{ij},\ j = 1, 2, \ldots, n\}$ be an independent random sample from the ith sub-sample, $i = 1, 2, \ldots, k$. The t-slippage problem is to test

H_0: $\theta_i = \theta$ for all i

vs. the slippage alternatives

$$H_J\colon \theta_j = \theta + \Delta, \quad j \in J$$
$$= \theta, \qquad j \notin J$$

where J is any subset of size t of the set $\{1, 2, \ldots, k\}$, $1 \le t \le k$.

If the ith sub-sample has a normal distribution with mean θ_i and a known standard deviation σ_i, a class of symmetric invariant Bayes tests for the t-slippage problem is [7]:

Accept H_0 if

$$\max_J (\bar{x}_J - \bar{x}) < C \tag{1}$$

accept H_I if

$$\max_J (\bar{x}_J - \bar{x}) \ge C$$

and $x_i \ge x_j$, $i \in I$, $j \notin I$, where $\bar{x} = (\sum_{i=1}^{k} \bar{x}_i)/k$ is the mean of the sub-sample means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, $\bar{x}_J = (\sum_{kJ} \bar{x}_i)/t$ is the pooled mean of the sub-samples $[i_1, i_2, \ldots, i_t\colon i_j \in J]$, and C is a real constant which characterizes the above class of Bayes rules.

The choice of C is arbitrary. We will choose C such that

$$P(\text{Accept } H_0 \mid H_0) = 1 - \alpha$$

which by Eqn. 1 becomes

$$P\left(\max_J [\bar{x}_J - \bar{x}] \le C\right) = 1 - \alpha \tag{2}$$

For the special case $k = 3$, $t = 2$, Eqn. 2 reduces to

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[1 - \Phi\left(\frac{\sqrt{n}\left[x\sqrt{6/n} - 6C\right]}{3}\right)\right]^3 e^{-x^2}\, dx = 1 - \alpha \tag{3}$$

The integral in Eqn. 3 can be numerically computed by using Hermite integration. Computations for C satisfying Eqn. 3 have been done for $n = 5$, 10 (1), 15, 20 and $\alpha = 0.10$, 0.05, 0.025, 0.01.

Table 1 shows the lower percentage points of the distribution of the test statistics $\max_J(\bar{x}_J - \bar{x})$

TABLE 1

Lower percentage points of the null distribution of $\max_J(\bar{x}_J - \bar{x})$

n	α							
	0.01	0.025	0.05	0.1	0.90	0.95	0.975	0.99
5	−0.2874	−0.2147	−0.1517	−0.0785	0.4619	0.5430	0.6144	0.6987
6	−0.2624	−0.1960	−0.1385	−0.0717	0.4217	0.4957	0.5609	0.7378
7	−0.2429	−0.1814	−0.1282	−0.0663	0.3904	0.4589	0.5193	0.5905
8	−0.2272	−0.1697	−0.1199	−0.0621	0.3652	0.4293	0.4857	0.5524
9	−0.2142	−0.1600	−0.1130	−0.0585	0.3443	0.4048	0.4580	0.5208
10	−0.2032	−0.1518	−0.1072	−0.0555	0.3266	0.3840	0.4345	0.4941
15	−0.1659	−0.1239	−0.0876	−0.0453	0.2667	0.3135	0.3548	0.4034
20	−0.1437	−0.1073	−0.0758	−0.0392	0.2310	0.2715	0.3072	0.3494

for the above values of k, t, and n. In order to use Eqn. 1, we need the upper percentage points of the distribution; these can be obtained from Table 1 using the symmetry of the null distribution.

The null distribution of the test statistic $\max_J(\bar{x}_J - \bar{x})$ is quite complicated for $t > 2$. Monte Carlo simulation can be used to compute the cut-off points for this case. Work in this direction is in progress, and will be reported upon completion.

EXAMPLES

In this section, Monte Carlo simulation is used to provide some illustrative examples of our multiple slippage testing procedure.

Example 1

In this example, $k = 3$, $n = 10$, $\theta_i = 0$, and $\sigma_i = 1$, $i = 1, 2, 3$. An IMSL subroutine was used to generate the following data:

$$\bar{x}_1 = 0.30469, \quad \bar{x}_2 = -0.09809, \quad \bar{x}_3 = -0.10785$$

The sorted observed means are:

$$\bar{y}_1 = -0.10785, \quad \bar{y}_2 = -0.09809, \quad \bar{x}_3 = 0.30469$$

$$\max_J \bar{x}_J = \frac{(\bar{y}_2 + \bar{y}_3)}{2} = 0.1033$$

and the pooled mean is

$$\bar{x} = \frac{\sum_{i-1}^{3} y_i}{3} = 0.03292$$

Hence the observed value of the test statistic is

$$\max_J (\bar{x}_J - \bar{x}) = 0.07038$$

For $\alpha = 0.10$, Table 1 gives $C = 0.3266$. Since the observed value of the test statistic (0.07038) is less than C, we do not reject H_0 at $\alpha = 0.10$. Notice that for this example, all the three population means are equal.

Example 2

In this example, $k = 3$, $t = 2$, $\sigma_i = 1$, and $\theta_1 = \theta_2 = 1$, $\theta = 0$. The generated sample means are:

$$\bar{x}_1 = 1.30070, \ \bar{x}_2 = 1.51989, \ \bar{x}_3 = 0.10179$$

from which the test statistic turns out to be

$$\max_J (\bar{x}_J - \bar{x}) = 0.4362$$

For $\alpha = 0.05$, Table 1 gives $C = 0.3840$. The null hypothesis of homogeneity is therefore rejected in favor of $H_{(1,2)}$: sub-samples 1 and 2 have larger means.

REFERENCES

1 F.E. Grubbs, Technometrics, 11 (1969) 1.
2 F.J. Anscombe, Technometrics, 2 (1960) 123.
3 V. Barnett and T. Lewis, Outliers in Statistical Data, Wiley, New York, 1978.
4 D.M. Hawkins, Identification of Outliers, Chapman and Hall, London, 1980.
5 F. Mosteller, Ann. Math. Stat., 19 (1948) 58.
6 S. Karlin and D.R. Truax, Ann. Math. Stat., 31 (1960) 296.
7 A.K. Singh, Can. J. Stat., 6 (1978) 201.

Analytica Chimica Acta, 277 (1993) 477–487
Elsevier Science Publishers B.V., Amsterdam

Simple regression and outlier detection using the median method

M.O. Moen, K.J. Griffin and A.H. Kalantar

Department of Chemistry, University of Alberta, Edmonton T6G 2G2 (Canada)

(Received 7th July 1992)

Abstract

Regression methods that are relatively unaffected by outlying data are necessary to provide unbiased estimates of the slope for data conforming to the straight line function. The non-parametric median method achieves this, using the median of the $n(n-1)/2$ possible pair-wise estimates of the slope, given n distinct data. Outliers in y are identified by comparing the (absolute values of the) y deviations of the individual data from the regression line with a robust measure of their dispersion, namely the median of all these absolute deviations (MedAD). Outliers so detected may be eliminated and a new regression line determined. Here the cutoff level for Q (the ratio of the absolute deviations to MedAD) is determined so as to maximize the number of outliers found and to minimize the number of non-outliers detected. This is done by simulating eight uniformly spaced data having outliers whose y-deviations are 3–10 times greater than those of the non-outlying points. By tracking the known outliers, we found $Q = 6$ to be a suitable cutoff. Elimination of known outliers reduces the variance of the slopes found using the median method. The efficiency of the method (with respect to least squares treatment) varied from 88% when no outliers were added to $> 600\%$ when one outlier (having 10 times the standard deviation) was added per data set. The slope of the remaining data should be determined by the median method rather than by least squares. We found that iterative treatment detected primarily non-outlying data and increased the variance of the resulting slopes. Non-outliers are more likely to be falsely identified as outliers if they happen to be central and not end points of the data set. Detection and elimination of outliers also works when x-axis errors are non-zero as long as the outliers' errors exceed the product of the slope and the x-axis errors. Recommendations for the experimenter are made and limitations of this study are delineated.

Keywords: Iterative elimination; Median method; Outliers

Regression of straight line data is commonly accomplished using least squares. If, however, the data set includes some outlying points, the least squares estimates of the slope and the intercept may be greatly affected and it is often difficult to detect outliers [1–3]. Consequently, regression procedures that are resistant to the presence of outliers have been studied and used in the past three decades. One such simple robust procedure, the non-parametric median method [4,5], is the major focus here.

Using this procedure for simple regression, each pair of data is used to calculate a slope. For n distinct data, there would be $n(n-1)/2$ slopes, the median of which is taken as the slope estimate, one that Sen showed to be unbiased [4]. (A similar procedure can also be applied to obtain an estimate of the intercept [6].) Because the slopes calculated using outlying points are usually found near the start or end of the ranked slopes, the median of such slopes is little affected and thus the median method is robust.

Correspondence to: A.H. Kalantar, Department of Chemistry, University of Alberta, Edmonton T6G 2G2 (Canada).

In spite of the median method's resistance to outliers, it has been suggested [7] that the procedure be extended to identify and then eliminate outliers before using it or another estimator to make a final estimate of the regression parameters. A study of this suggestion constitutes a major part of this work.

Before detailing and examining this modified regression procedure, however, it is useful to describe the analogous but simpler one-dimensional case where such a procedure was found to be successful and efficient [8]. Using that procedure, the central tendency of some outlier-containing data is described by the very robust median. Then the deviations of each datum from the median are calculated. Their absolute values are ranked and the median of these absolute values is determined. This is called the median (of the) absolute deviations or MedAD. The absolute values of the deviations are then compared to MedAD by forming their quotient, Q. Whenever this quotient exceeds 5.2, the corresponding datum is eliminated [8]. Finally, the mean of the remaining data is used as the estimate of central tendency for the data set. (It is also possible to repeat this procedure, to eliminate more data, iteratively [9].) It was found that the efficiency of this procedure (involving a Huber-type skipped mean and called X84 [10]) is ca. 90% (for 8 data), whereas the efficiency of simply using the median of the untrimmed data set is only 64% compared to the mean [2], for data with normally distributed errors having no outliers.

The robust procedure suggested [7] for simple regression is similar. First, the median method is applied to find the slope for n data. Then, using the robust slope estimate, n values of the intercept are calculated, ranked and their median used for the robust estimate of this parameter [7,11]. Then the y-deviations of the n data are calculated and their MedAD found. The quotients, Q (the absolute value of the y-deviations divided by MedAD), are formed and compared to a cutoff value. A datum is eliminated if its Q exceeds the cutoff value ($Q = 3$ in Ref. 7) and the remaining data are treated again with the median method. This procedure is repeated until no further outliers are so identified. The final estimates of the parameters are determined by re-applying the median method to the remaining data. Rock and Duffy [12] have published a program to carry out this procedure. (Rousseeuw and Leroy [3] recommend a related procedure using the least median of squares method [13] first and then least squares, but there is only one outlier rejection step.)

This procedure, in light of the one-dimensional case, appeared promising enough to warrant investigation of its efficiency. We were also concerned about the cutoff for Q because a simulated data set [14] having no outliers yielded two outliers using this procedure [7] and because its cutoff choice was little tested. Thus the specific questions suggested are: What cutoff value of Q should be used? What is the efficiency of the method? Should points determined to be outliers be eliminated iteratively? Should the remaining data be treated by least squares or by the median method? Do biases result? Do the two procedures for obtaining the intercept [6,11] yield significantly different results?

The answers to these and related matters have been obtained by repeatedly simulating data for a few cases and monitoring the analyses carefully when outliers were both omitted and added. This was done only for errors that were independent, identically and normally distributed and for outliers (in only y) that differed from non-outlying data only in their greater variances. After the simulations and results are described, we make recommendations about using this procedure and note the limitations of this narrowly focussed study so that the results will not be inadvertantly misapplied.

SIMULATIONS

Data sets were simulated by adding pseudo-random numbers to a set of points following exactly the straight line relation $Y_i = a + bX_i$. Eight equally spaced values of X_i (ranging over 10 units) were used; these were 0.6667, 2.0952, $\cdots$, 10.6667. For the cases detailed below, $a = 0.7$ and $b = 1$. This gives, for example, $Y_3 = 4.2238$. Data sets consisting of eight points

are commonly found and simulated [7,15,16]. Such a data set provides enough data to treat after outliers are discarded, while minimizing computing time, which increases as n^2. Eight is also the minimum number that permits two outliers to be handled by the median method without exceeding its breakdown point [17].

To the exact Y_i were added computer-generated pseudorandom numbers (ϵ_i) that were normally distributed about zero and had the same variance (σ^2), typically 10^{-4}. These provided the 'noisy' data (y_i) with homoskedastic errors, ϵ_i: $N(0,\sigma^2)$ via $y_i = Y_i + \epsilon_i = a + bX_i + \epsilon_i$ as noted elsewhere [18]. Our test of several million such ϵ_i, $N(0,\sigma^2)$ showed very close adherence to the expected distribution, to $\pm 4\sigma$, as did the distribution of consecutive values of these numbers with the same sign. We also checked and found no evidence whatsoever for any serial correlation, for periods of 1, 2, 5, 8, 12 or 16. We checked the pseudorandom generator further by using a second multiplicative congruential generator that has also been widely tested. Both generators yielded indistinguishable results. Thus the ϵ_i were identically and independently normally distributed for our purposes. Finally the X_i were also converted to x_i by similarly generated but smaller ϵ_i. For the x errors, $\sigma^2 = 10^{-10}$, making the x_i essentially error-free relative to the y_i.

The outliers were introduced in only the y_i. They differed from the non-outliers in only one respect: the variance of their ϵ_i was made larger. The outliers' errors were $N(0,c\sigma^2)$, with c usually 9 or 100. Thus their inclusion yielded a heavy-tailed distribution. Such outlying points can thus represent some large errors or disparate samples of other data sets (as in [7]). These were introduced regularly so that each of the eight y_i was made outlying an equal number of times. The fractions of outliers used were 1/8, 1/4 and 3/8. For example, for 1008 simulated sets of data and 3/8 outliers, each of the (56) permutations of exactly three outliers per data set was present 18 times and each y_i was drawn from the broader distribution 378 times. Of course a few of these $N(0,c\sigma^2)$ outliers were less deviant than the $N(0,\sigma^2)$ non-outliers because their pseudorandom ϵ_i happened to be smaller. About 5% of these forced outliers were, in fact, closer to zero than some larger deliberate non-outliers, for c = 100.

As the simulated data sets were prepared and analyzed, several indices, etc., were accumulated to monitor whether or not, for example, the known outliers were correctly detected. Exactly what was monitored will become clear as the results are presented and discussed. The slopes, intercepts, MedAD, etc., were also stored. These enable standard deviations of the slopes, etc., to be calculated, so that the efficiencies (the ratios of the variances) of various methods of analysis may be compared quantitatively. These comparisons are made more valid when these standard deviations are better determined. This occurs when the number of simulated data sets rises. In general, each case examined here involved (at least) 4032 simulated data sets. This means the relative standard deviation of our calculated standard deviations (for example, of the slope) are 1.11% [19] and so the relative standard deviations of the efficiencies reported here are no greater than 3.17%. These efficiencies are usually referenced to the results obtained by least squares analysis of the simulated data.

Because of the errors in the x_i (however small) in addition to the errors in the y_i, we used the treatment of Williamson [20] to extract the parameters without the bias that ordinary least squares yields. Williamson's treatment requires estimated variances for the weights for x_i and y_i, variances we specified and used to generate the 'noisy' data. The same variances (σ^2) were used for each datum, whether made to be outlying ($c\sigma^2$) or not (σ^2).

RESULTS AND DISCUSSION

Simulated data with no deliberately introduced outliers

Although the major thrust of this study is to determine how well these procedures distinguish between outlying and non-outlying data, we begin with the case where *no* outliers have been introduced. This is done because it is a suitable case

for illustrating and explaining the simulation, the information monitored and the results. It also provides a basis for comparison and permits a number of consistency checks. Finally, it happens that several outlier-containing cases yield results that are similar to the case with no outliers. In what follows, we will examine the slopes and intercepts determined and then consider the MedAD and the Q (quotients of the y-deviations with respect to MedAD). These Q will be used for distinguishing between outliers and non-outliers. Patterns of the Q within each data set are found and explained and then the distribution of the Q for all data sets is compared as a prelude to considering the most appropriate cutoff value for Q.

Slopes, b. As noted in the introduction, n data yield $n(n-1)/2$ pair-wise slopes (28 for our $n = 8$ case) which are ranked. Thus, the median slope is the average of the fourteenth and fifteenth ranked slopes. The median of these 28 slopes is predicted [4] and found to be unbiased, as is the slope determined by least squares. Their standard deviations [$\sigma(b)$] differ: $\sigma(b)$ for the median method is larger than it is for least squares by 7%. That is, the efficiency of the median method is 88($\pm$0.5)% that of getting the slope via least squares when outliers are absent when *all* the ϵ_i are $N(0,\sigma^2)$, for 8 data. This same efficiency was found earlier [21]. Sen has shown [4] that the scale estimate, $\sigma(b)$, for the median method may be determined from the ranked slopes. As indicated elsewhere [22] the formulae [4] and application [7] of this procedure are incorrect. The range encompassed by the seventh and twenty-second ranked slopes is expected and is found to include the correct slope in 94.6($\pm$0.5)% of all simulations.

Intercepts, a. As with the slopes, a pair of data points can yield an intercept and one could likewise use the median of the 28 possible pair-wise intercepts for a robust estimator of the intercept [6]. A second method uses the slope found via the median method, together with each of the 8 data points to yield 8 estimates of the intercept [11]. The median of these is taken as the estimate of the intercept. Both of these two robust methods of estimating intercepts are equivalent with respect to both the means and the $\sigma(a)$ found. As a consequence we have routinely used the second method, involving the median-determined slope and the median of the eight intercepts estimated therefrom. The efficiency of either procedure is about 83% for this case, with respect to the intercept determined by least squares when outliers are absent.

MedAD. Once the intercept is determined, one may examine the differences between each of the eight intercept estimates or, equivalently, the differences between the observed y_i and the calculated $\hat{y}_i$ ($= a + bx_i$). The median of the absolute values of these differences is the MedAD and it serves as an estimator of the spread of the data in the y direction. Unlike σ, MedAD is robust, i.e., it is relatively insensitive to one or two very deviant (outlying) y_i. Its average value reflects this as it is smaller than the standard deviation of the y-deviations, even when no outliers are present.

Q. In an effort to detect possible outliers, we compare each of the eight y-deviations of a given data set to that set's MedAD via the quotient, $Q_i = (|\, y_i \text{ deviation} \,|)/\text{MedAD}$. For each data set, the eight unranked Q_i were accumulated to provide the means and spreads for each of Q_1, $Q_2, \ldots, Q_8$. A clear and symmetric pattern emerged: on average, the first and last data points are closest to the median method's regression line while the central points (4 and 5) are the farthest from it. Put another way, points 1 and 8 have more influence on the regression line than points 4 and 5. (For the least squares line, these influences are even more different.) This was examined in more detail by keeping track of which data points ultimately gave rise to the median slopes and intercepts. This means we counted the number of times the pairs (1 and 2, 1 and 3, ... and 1 and 8, etc.) share in the determination of b, i.e., occur in the fourteenth or fifteenth ranked slopes. We found that, of the 28 pairs of slope-determining data, those most frequently found to contribute to the median slope were those coming from the most distant pairs of data [(1,8), (1,7), (2,8), etc.]. We also determined the standard deviations of each of the *un*ranked slopes. These were consistent: the closer together

TABLE 1

Number of times data points are most deviant before and after fitting by the median method when no outliers have been included for 8064 sets of 8 equally spaced data

Known before		Observed after fit			
		1,8	2,7	3,6	4,5
1,8	2012	760	395	450	407
2,7	2003	253	1127	286	337
3,6	2081	213	196	1438	234
4,5	1968	103	123	203	1539
Totals	8064	1329	1841	2377	2517

were the data used to calculate a slope, the larger was that slope's standard deviation.

One may see this algebraically by considering the expression for the error in the slope for the pair (x_i,y_i) and (x_j,y_j):

$$\frac{y_i-y_j}{x_i-x_j}-\frac{Y_i-Y_j}{X_i-X_j}=\frac{y_e-(\text{slope})x_e}{X_i-X_j+x_e}$$

where, e.g., y_e is the difference in the errors of the y_i or $\epsilon_i-\epsilon_j$. If the x_i are error-free or controlled in the Berkson sense [23], then the error in the slope is simply $y_e/(X_i-X_j)$ for the i,j pair of data and so it is merely inversely proportional to the x-separation of the points.

The unequal influences of the central and end points of the data sets have other effects. Suppose we determine which y_i is the most deviant from the true line (before any fitting) and then monitor whether it is still the most deviant with respect to the robust regression line found for this data set (after fitting). This has been done and the results are presented in Table 1. The entires of the second column show, as expected, that each y_i is equally likely to be the most deviant from the true line *before* fitting. The last row shows the bias in the deviations in the y direction *after* fitting. The diagonal of the table shows that generally (though hardly exclusively) the most deviant y_i can be the same before and after the fit. Should some points most deviant from the fitted line be declared outliers, then Table 1 suggests that central points will be identified as outliers nearly twice as frequently as will end points. Thus there may be a bias associated with outlier direction, one that arises from the unequal influences of these points on the regression line when using this median method. For the double run (8064 simulations) used for Table 1, this bias is 1.94 ± 0.09 ($= 2517 \div 1329$, normalized by $2012 \div 1968$). However, there is no bias in the average obtained for the 8064 values of either the slope or the intercept.

Frequency distribution of Q. Q varies within a given data set because of the varying ϵ_i and the deviations of the y_i from the regression line. Q also varies between data sets because the MedAD vary. In order to choose a suitable Q beyond which a datum would be declared an outlier, we determined the frequencies presented in Table 2. The entries, for each range of Q, represent the number of times Q fell within the range (out of 8064 sets of 8 values of Q). Row 2 shows the situation after the first fit and before any cut or elimination of data has been made. Row 5 gives the cumulative count of the Q. Thus, in the

TABLE 2

Frequency distribution of Q (y deviations/MedAD), for the case of Table 1 [a]

1	0	2	3	4	6	8	9	10	15	20	> 20
2	50649	6158	3038	2725	1001	268	146	356	84	87	
3	2481	198	163	188	115	43	39	122	44	62	
4	726	80	37	37	19	14	5	8	5	22	
5	64512	13863	7705	4667	1942	941	673	527	171	87	
	0	2	3	4	6	8	9	10	15	20	> 20

[a] Row 1. Range of Q values. Row 2. Number of points within Q ranges after first fit, before any data have been eliminated. Row 3. Number of points within Q ranges after one cut at $Q = 9$ and after a second fitting. Row 4. Number of additional points within Q ranges that are subsequently eliminated for cases where iterative elimination of data has taken place. Row 5. Cumulative totals of number of data in these Q ranges for first fit (row 2), before elimination of any data.

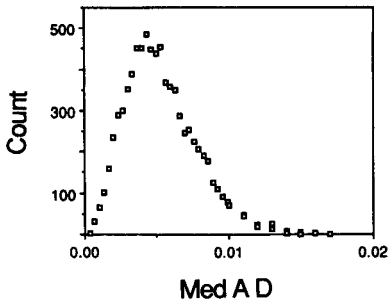


Fig. 1. Frequency distribution of MedAD for 8064 sets of 8 uniformly spaced $N(0,\sigma^2)$ data when no outliers have been introduced. Here the median (mean) MedAD is 0.0048 (0.0052).

penultimate column of row 2, '84' represents the number of times Q exceeded 15 but was ≤ 20 after the first fit. In row 5, the entry '527' represents the total number of times Q exceeded 10. Thus we see that Q exceeds three 7705 (or about 12%) times. As this table and other runs show, the demarcating Q should be about 6 if one regards about 3% of these data as sufficiently deviant to warrant possible identification as outliers. [This Q value is reminiscent of the cutoff ($Q = 5.2$) used in the one-dimensional study [8], for the X84 estimator.] However this cutoff is only a preliminary guess: we must first see what happens when outliers are actually included.

Most striking is the fact that several Q values are very large (the largest is > 120 for the double run giving rise to Tables 1 and 2). In terms of the σ used to generate the ϵ_i, we would expect $Q < 5\sigma$ if only the ϵ_i affected Q. Evidently the MedAD are very variable and can be quite small: that is, the Q seem to be MedAD-driven. Figure 1 displays the distribution of the MedAD for the Tables 1 and 2. It demonstrates that there are indeed very many quite small MedAD values. To see if the Q are related to their respective MedAD, we collected both the MedAD and the two largest values of Q for each of 4032 data sets. The indices of the 300 smallest MedAD were then compared with the indices of the 300 largest Q (each index shows which of the 4032 sets of data was analyzed). For the two collections of 300 such indices, one expects only 20–25 to match if the largest Q are unrelated to the smallest MedAD. The typical overlap (same indices) found is > 200 (and 15–20 when considering only the most extreme 30 of the 4032 runs). Moreover, about a quarter of the largest Q are the second largest Q from the same data sets. Clearly then it is a small MedAD for a given data set that simultaneously inflates the two largest Q collected from that data set and, more generally, that the large Q really are driven by the small MedAD.

Consequences of rejection of data. If the data most deviant from the robust regression line are left out, a new line can be determined, using least squares or the median method. For the present case where no outliers are present, the mean of the new slopes for the trimmed data sets is neither consistently closer to the true slope nor is it significantly different from the mean of the original slopes. However, their $\sigma(b)$ are always *larger* (as there are now fewer data), showing that such a procedure is inefficient, at least in the absence of outliers. Deviations from the second fitted line, using the median method, can also be compared and further data can sometimes be rejected. Such an iterative rejection procedure was examined in which we stopped when either not further outliers were found or when a total of ≤ 4 data had been rejected in ≤ 4 cycles, whichever occurred first. (If a fifth datum is eliminated, then a sixth datum will necessarily be subsequently eliminated. When the cut was made at $Q = 2$, it was found [7] that all but two data were eliminated as outliers. Naturally if this were the case, it would be wrong to use this low breakdown median method.) Again, $\sigma(b)$ always increased and the mean of the new slopes was not significantly different from the mean of slopes for data sets where some data had been eliminated.

Row 3 of Table 2 shows what happens, for $Q = 9$, after the first cut [673 ($= 146 + 356 + 84 + 87$) or about 1% of the data have been rejected as a result of the first fit]. In these data sets, the mean MedAD is lowered by a factor of about 4 and so Q increases, yielding a further 267 ($= 39 + 122 + 44 + 62$) data whose Q now also exceed 9. Thus while none of these data was an outlier, an additional 267 were so labelled, largely be-

cause of the drop in the MedAD in ≤ 673 of the data sets. This suggests that if we examine data having real outliers and if we correctly identify and reject such outliers, then subsequent application of this method will involve lower MedAD. Thus some newly declared outliers may well be non-outliers whose Q values have been inflated by the lower MedAD. This possibility will be examined below. Row 4 in Table 2 shows how many more points (beyond those already 'found') are eventually deemed to be outlying in subsequent iterations (40 more, for $Q = 9$).

To review the findings when no outliers are present, we note that we have the efficiencies of finding the regression parameters, MedAD is variable and frequently small (Fig. 1), the large Q are MedAD-driven and the frequency distribution of the Q suggests that a cutoff rather greater than $Q = 3$ may be more appropriate for detecting outliers. It also seems that iterative elimination of outliers may be inefficient. Finally, we found a significant bias in the method's attribution of outlying character, according to whether the datum in question is an end point or a central point (Table 1).

Simulated data with outliers deliberately included

Including outliers among the data allows us to determine how robust the median method is, the extent to which this procedure can detect known outliers and also permits us to find a suitable cutoff for Q. Such a cutoff will be a compromise between correctly finding as many true outliers as possible while minimizing a Type II error: the false detection of non-outliers. Unless stated otherwise, we concentrate on the relatively extreme [2] case of 25% outliers [$N(0,100\sigma^2)$] in the presence of 75% non-outliers [$N(0,\sigma^2)$]. This is just below the finite sample breakdown point for the median method (29% outliers), beyond which a bias in the regression parameters for a data set would be unbounded as the outlier's variance increased indefinitely [17].

Comparisons with non-outlier cases. Many of the results with outliers are similar to the results for non-outliers. The median method continues to provide unbiased parameters, but their standard deviations rise (that is, the median method is not totally resistant to outliers). Once again, the range encompassed by the seventh and twenty-second ranked slopes is found to include the correct slope in 94.6% of all simulations even though the errors belong to two normal distributions having very unequal variances (σ^2 and $100\sigma^2$). Intercepts determined by either of the procedures [6,11] described earlier are again indistinguishable with respect to mean or standard deviation. Of course, the least squares procedure, which is sensitive to the presence of outliers, yields a greater range of regression parameters. In fact least squares becomes much less efficient

TABLE 3

Frequency distribution of Q for 8064 sets of 8 equally spaced data points of which exactly two are meant to be outliers: 25% $N(0,100\sigma^2)$ [a]

1	0	2	3	4	6	8	9	10	15	20	> 20	
2	3175	1136	1033	1880	1535	692	646	2242	1353	2436		OUT
3	1509	419	351	561	310	105	81	211	70	115		OUT
4	238	25	15	15	3	1	2	1	2	5		OUT
5	16128	12953	11817	10784	8904	7369	6677	6031	3789	2436		OUT
	0	2	3	4	6	8	9	10	15	20	> 20	
6	43657	2623	1011	720	194	38	42	64	26	9		
7	29775	2889	1505	1511	707	220	172	374	146	236		
8	7258	749	397	452	84	21	25	99	48	181		
9	48384	4727	2104	1093	373	179	141	99	35	9		
	0	2	3	4	6	8	9	10	15	20	> 20	

[a] Rows 1–5. Results for outlying data. See rows 1–5 of Table 2. Rows 6–9. Same as for rows 2–5, but for the 75% of data that are not made deliberately outlying. Rows 3,7, result after just one cut, at $Q = 6$. Rows 4,8, result for all subsequent cuts at $Q = 6$.

as the fraction and/or variance of the outliers is increased (it is $< 20\%$, compared to the median method).

The MedAD values also increase with the number and variance of the added outliers, but their distribution continues to look like that shown in Fig. 1. Their mean (median) is 0.0088 (0.0079). The pattern and symmetry of the mean Q at each datum is also unchanged, but their mean values are more than doubled. As the variance of the outliers increases, the counts in the analogue of Table 1 become much more concentrated in the four diagonal elements. That is, the most deviant points before fitting are much more likely to remain the most deviant after fitting if they have larger variances (are outliers). Thus the differences in detection of end points vs. central points are reduced [the bias is < 1.2 (± 0.023) for this 25% $N(0,100\sigma^2)$ case].

Finding a cutoff for Q for outlier-containing data sets. The goal of distinguishing between outliers and non-outliers (and therefore establishing a suitable cutoff) can be accomplished if the outliers and non-outliers are tracked. This is possible with the computer and is greatly simplified by the power of APL. That such a monitoring scheme is unambiguous is shown in Table 3. The frequency counts in the various Q ranges are displayed as in Table 2: rows 2–5 are for the 25% of the points made outlying with $100\sigma^2$ and the last four rows (6–9) of counts are for the non-outliers. Rows 2 and 6 show the distribution of the Q values after the first fit by the median method while rows 5 and 9 give the corresponding cumulative counts. Comparison of row 6 of Table 3 with row 2 of Table 2 shows that many of the non-outliers have shifted to lower Q in the presence of 25% outliers because the MedAD are larger. On the other hand, row 2 shows that many of the outliers are concentrated at high Q (the largest exceeds 500), in spite of the greater MedAD, and so dominate the counts at large Q. Thus 1353 outliers (row 2) and only 26 non-outliers (row 6) have Q in the 15–20 range and 6031 (row 5) of 6130 (or 98%) of the Q exceeding 10 are outliers. These results, for this relatively extreme case, yield a cutoff dependent upon only specifying some criterion. One criterion is to extend the cutoff to the next lower Q only if the number of additional outliers detected is no less than the number of additional non-outliers falsely detected. On this basis, admittedly arbitrary, we found $Q = 3.6$ to be appropriate for this 25% $N(0,100\sigma^2)$ and $n = 8$ case. This would identify 69% of the outliers and 3% of the non-outliers as outlying. Clearly 31% of the points deliberately made outlying have low Q values (because of the larger MedAD) and so have not been labelled outliers. Of course the experimenter would not usually know either the number or the variance of any possible outliers, let alone which data were outliers. This knowledge and the simulations help us to understand the procedure so as to find a reasonable compromise cutoff that could be suitable for a variety of cases.

Is a second cut advisable? In the example being used here, it happens that we deliberately used a higher cut —at $Q = 6$. This finds 56% of the outliers, while incorrectly identifying only 0.8% of the non-outliers. Thus 44% of the outliers remain. Therefore, we will re-fit the remaining data after making the first cut at $Q = 6$. This reduces the average MedAD to that of Fig. 1. As in row 3 of Table 2, rows 3 and 7 of Table 3 give the counts of additional Q found in these ranges after the first cut and the second fit. The lower MedAD, together with new y-deviations, yield more Q that now exceed 6 for the first time. Of these, merely a third (892 of 892 + 1855, the last 6 entries of rows 3 and 7) are known to be real outliers at any but the lowest Q intervals. Using the criterion given above, we would advise against a second cut. Had the first cut been made at the more appropriate $Q = 3.6$, we would have found that after the second fit fewer than 10% of any newly fit points whose Q now exceed 3.6 (or even 9) would be genuine outliers. One could also consider making a second cut at a *greater* cutoff value of Q, to offset the effect of a smaller MedAD. We found this to be generally fruitless even though we, unlike the experimenter, knew the extent of outliers present. A second cut that is *lower* is also possible. This could be useful if, for example, the first cut had been at $Q = 20$ and had affected perhaps only 1% of the data sets. The absurdity of 'finding' outliers that are mostly

non-outliers worsens for all subsequent cuts. As rows 4 and 8 of Table 3 show for a $Q = 6$ cut, about 97% of those cut after the second cut would actually be non-outliers. These represent an upper limit (5.8%) on the number of data sets identified as having more than two outliers.

Thus the separate frequency counts —for both tracked outliers and non-outliers— show that iterative outlier rejection and refitting procedures are usually highly misleading. Rejection beyond the first round involves the rejection of predominantly good data, not outliers. Such rejection involves a similar but less pronounced midpoint vs. end point bias noted earlier and leads to inflated, rather than reduced, standard deviations of the regression parameters sought. In general the mean of the estimated slopes is not significantly affected.

Considerations associated with the first cut. Should the data remaining after one cut be analyzed using the median method or least squares? For the relatively extreme case focussed upon thus far, it happens that $\sigma(b)$ decreases appreciably for both the median method and least squares analyses but the median method remains the more efficient after a 3.6 cut. However, as the fraction and especially the variance of the outliers decreases, higher cuts become more common. A choice for a cut depends upon the collection of cases examined. In our judgement, a suitable compromise appears to be a 6 cut. Even if this is not an optimal cut, any difference in efficiency of analyses by the median method vs. least squares is minimal (10–20%) unless we are near the 25% outlier or $100\sigma^2$ contamination level, in which case the median method is necessary. The situation is the same with respect to data rejection: this appears to make a difference only for the more extreme contamination levels; otherwise efficiency losses or gains are minimal.

Other cases have been examined, e.g., for outliers having smaller variances and/or for 1 or 3 outliers. (For 3 outliers added, it is strictly proper to use only methods [13,23] having higher breakdown points, especially for larger variances. However, the efficiencies of these more robust methods are substantially lower.) A few results for some of the other cases are summarized in Fig. 2,

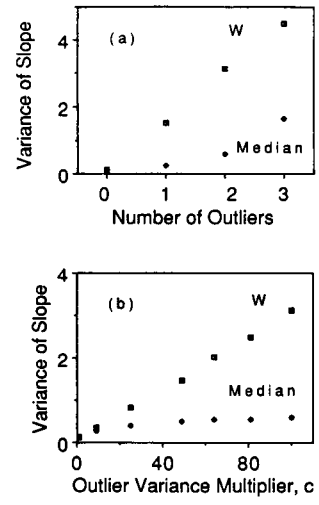


Fig. 2. Variances of the slopes (all multiplied by 10^5) for estimation by least squares and the median method on all the data as a function of (a) the number of outliers (with variance $= 100\sigma^2$) per set of 8 uniformly spaced data and (b) the variance of the outliers for 25% $N(0,c\sigma^2)$ outliers. Each plotted datum is for 4032 runs; in all cases the variances of the ϵ_i are 10^{-6} for x and 10^{-4} for the non-outlying y and the slope is 1. W refers to least squares treatment in Ref. 20.

which contrasts the effects of outliers and their variance upon the slope's variance, using both least squares and the median method on all the data. Results for intercepts have been omitted, partly because they also depend upon the extent of extrapolation involved. Clearly, even without detection and elimination of outliers, the median method is very much more robust. The central results of this sub-section are the establishment of an approximate cutoff ($Q \approx 6$) so as to identify outliers and the confirmation that iterative elimination of data later called outlying is unwise because their majority is, in fact, not outlying. Tracking the outliers (see Table 3) has been invaluable for this study.

Although we have been concerned with eliminating detected outliers, it should also be noted that reliable detection alone can also be valuable. Their cause(s) can be investigated so as to improve controls or measurements or to reconsider the model being used for fitting. It has also been pointed out [2] that identification of outliers has led to new patented processes.

Finally we note some similarities and differences of this detect-and-discard procedure for the central tendency (X84) and regression cases. In the former, use of the low-efficiency, high-breakdown median yields, after a single elimination step, a very efficient (ca. 90%) estimate of the central tendency. For the regression case, there is a big improvement in the efficiency only for cases in which the variance of outliers is fairly high. For both cases, use of the mean or least squares, respectively, is very seriously affected by the presence of outliers.

RECOMMENDATIONS

These simulations show that the procedure of using the median method on straight line data and then forming the quotient (Q) of the y-deviations and their MedAD succeeds in identifying outliers in the cases to which we have restricted our discussion. A cut at $Q = 6$ appears to be a reasonable compromise between eliminating too few outliers or too many non-outliers, for 8 data. Finally, the regression parameters for the remaining data should also be determined by using the median method after only one cut. Further deletions of data are self-defeating and unwise because many more non-outliers are eliminated than outliers and because $\sigma(b)$ and $\sigma(a)$ increase.

This procedure will be only 80–90% efficient if the data are relatively free of outliers. Thus this cost (efficiency loss) is the 'insurance premium' for protection against outliers, a protection that least squares cannot provide (Fig. 2). If more than 10–15% of the data could have an appreciable variance, this procedure is more efficient than not eliminating the outliers detected or using least squares after deleting some data. For still more protection, one may want to consider the repeated median [23] or the least median of squares [13] method, whose finite sample breakdown points are 50%, but whose efficiencies are even lower [3] than that of the median method [4]. Of course the experimenter does not know the fraction, extent or nature of outliers present. Ultimately, however, it is the experimenter's experience and perception of the level of outliers that must guide a choice of method and a particular cutoff for Q. This study shows the minimum that must be considered before adopting any relatively untested suggestion.

These recommendations apply to the simulations considered here. These conform to the usual assumptions: independent errors drawn from the same normal distribution for the non-outliers and from a single broader normal distribution for the outliers thus yielding heavy tails rather than skewed combined distributions. While these assumptions may well be unrealistic, they often correspond to the assumptions implicit in the regression techniques as they are usually applied, such as least squares. These recommendations apply to equally spaced data; Sen showed [4] that, for cases without outliers, the efficiency associated with other spacings was less by the factor ρ^2, where ρ is the correlation coefficient between the x_i and the i. Errors in x are permissible if they are small enough to leave the x_i properly ordered. In addition, the slope times the error in x must not be too large compared to the outliers' errors (in y), otherwise outliers in y cannot be successfully detected using MedAD and Q.

If one must treat many data sets, it would be prudent to keep track of the position of the data being detected as outlying by the method. This is because any noticeable bias in the deletion of more central than end points may mean that one is really deleting relatively more non-outliers (Table 1).

For other untested cases, specific simulations are necessary. Our study shows that such simulations should include monitoring of outliers and non-outliers separately. Even with the ease of programming permitted by high level languages such as APL, the simulation requires effort and care. This is worthwhile, we believe, because of the benefits provided. As indicated, these include an appreciation of the levels and consequences of Type I and Type II errors, the efficiency of the method of analysis and the variances of the parameters sought. Any simulations may, as ours did, also yield findings that are unexpected and so lead to greater care or a better understanding of the procedure and the variables. Such appreciations will permit a more realistic assessment of

the experimental results. Such simulations would be limited only by the experimenter's knowledge of their systems, a knowledge that would surely lead to simulations that mirror their systems more closely than the examples above might.

LIMITATIONS

This study has been narrowly focussed and idealized. Although this is appropriate for a first, exploratory survey, it is unlikely to serve as more than a guide to the experimentalist. Thus the close of the Recommendations section noted the value of simulations based on one's own system. Here we will emphasize these points by restating some assumptions in a way that brings out what has not been considered, lest one inadvertantly infer otherwise.

The median method can handle other error distributions [5] but this work has used only normally distributed errors. Error structures other than simple errors (e.g., equal relative errors) have also been ignored. The simulation of errors has been likewise limited. So far, only those of scale and not in mean have been considered and then only along the y-axis and having the same error structure as for the non-outlying data. Our results are for a single case: exactly the same fraction of outliers for every data set, rather than the more realistic cases permitting some distribution of 1, 2 and 3 outliers such that the *average* yields the desired fraction of outlying data. [We have also examined a few cases involving outliers in location only, e.g., their mean is displaced from zero by ca. 10σ while their variance is the same ($c = 1$). Such outliers seem to be much more easily distinguished from non-outliers and appear to present few problems.]

It is known that the cutoff and the efficiency are affected by the number of data [8]; some significant differences might appear for rather smaller or much larger data sets. The small variances used here for the ϵ_i meant that the data remained monotonic. If the ϵ_i are increased sufficiently for at least the outliers in y, the median method yields a biased slope. Such cases were also ignored here.

We thank Computing Services for continued support, the Province of Alberta for a STEP grant and the Central Research Fund of the University of Alberta for partial financial support.

REFERENCES

1 V. Barnett and T. Lewis, Outliers in Statistical Data, Wiley, New York, 2nd edn., 1984.
2 F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, Robust Statistics, Wiley, New York, 1986, Chap. 1.
3 P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987, Chaps. 1, 2 and 4.
4 P.K. Sen, J. Am. Statist. Assoc., 63 (1968) 1379.
5 M. Hollander and D.A. Wolfe, Non-Parametric Statistical Methods, Wiley, New York, 1973, Chap. 9.
6 E.J. Dietz, Am. Statist., 43 (1989) 35.
7 R.G. Vugrinovich, Math. Geol., 13 (1981) 443.
8 M. Schweingruber, Das Monte Carlo Verhalten einiger Verwerfungsregeln, Thesis, ETH, Zürich, 1980.
9 F.J. Anscombe, Technometrics, 2 (1960) 123.
10 D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers and J.W. Tukey, Robust Estimates of Location, Wiley, New York, 1971.
11 J.S. Maritz, Austral. J. Statist., 21 (1979) 30.
12 N.M.S. Rock and T.R. Duffy, Comput. Geosci., 12 (1986) 807.
13 P.J. Rousseeuw, J. Am. Statist. Assoc., 79 (1984) 871.
14 C. Brooks, S.R. Hart and I. Wendt, Rev. Geophys. Space Phys., 10 (1972) 551.
15 L. Endrenyi and H.Y. Tang, Comput. Biomed. Res., 13 (1980) 430.
16 G.L. Atkins, Comput. Biol. Med., 12 (1982) 201.
17 D.L. Donoho and P.J. Huber, in P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr. (Eds.), A Festschrift for E.L. Lehmann, Wadsworth, Belmont, CA, 1983, p. 157.
18 A.H. Kalantar, J. Chem. Educ., 68 (1991) 368.
19 J. Topping, Errors of Observation and Their Treatment, Chapman and Hall, London, 3rd edn., 1965, p. 89.
20 J.H. Williamson, Can. J. Phys., 46 (1968) 1845.
21 A.H. Kalantar, Comput. Biol. Med., 17 (1987) 209.
22 A.H. Kalantar, Math. Geol., 22 (1990) 145.
23 J. Berkson, J. Am. Statist. Assoc., 45 (1950) 164; A.F. Siegel, Biometrika, 69 (1982) 242.

Analytica Chimica Acta, 277 (1993) 489–494
Elsevier Science Publishers B.V., Amsterdam

Outlier detection by robust alternating regression

Åsmund Ukkelberg and Odd S. Borgen

Division of Physical Chemistry, Norwegian Institute of Technology, University of Trondheim, N-7034 Trondheim (Norway)

(Received 3rd September 1992)

Abstract

The sum of least-squares regression method is normally used when principal components are extracted from a data matrix. This may result in a misleading set of principal components if outliers are present in the data set, in terms of both the number of components and their direction in vector space. Therefore, a robust alternating regression method is proposed. This method can be used to detect and correct or eliminate outliers.

Keywords: Alternating regression; Outlier detection

In principal component analysis, one wishes to find a bilinear model where the data matrix $\boldsymbol{R}$ is the product of a score matrix $\boldsymbol{S}$ and a loading matrix $\boldsymbol{L}$

$$\boldsymbol{R} = \boldsymbol{S} \cdot \boldsymbol{L}^{\mathrm{T}} + \boldsymbol{E} \quad (1)$$

where $\boldsymbol{E}$ represents the model residuals. This decomposition can be seen as a coordinate transformation where the orthogonal columns of the loading matrix denote the direction of the new axes and the likewise orthogonal columns of the score matrix represent projections of the data points on to these. These factors or bilinear components will describe the variance of the data matrix in descending order. The pseudo-rank of the data set denotes the number of components necessary to reproduce $\boldsymbol{R}$ with the necessary accuracy while excluding noise optimally. The pseudo-rank, score and loading matrices of the decomposition are all sensitive to outliers. Outliers in a data set represent data elements that are abnormal compared with the majority of the data. An outlier may belong to a different class of data, thus providing valuable information, or be caused by some sort of error. See Martens and Næs [1] for a further discussion. They used a combination of leverages and model residuals as outlier criteria. These criteria are based on the assumption that the least-squares model obtained gives a good representation of the majority of the data. Rousseeuw and Leroy [2] gave examples for which this assumption is wrong.

Correspondence to: O.S. Borgen, Division of Physical Chemistry, Norwegian Institute of Technology, University of Trondheim, N-7034 Trondheim (Norway).

THE NIPALS ALGORITHM AND ALTERNATING REGRESSION

The NIPALS algorithm (originally due to Wold [3] and made popular in chemometrics by Wold et al. [4]) is a useful tool for extracting principal components from a data matrix. An outline of the algorithm is given below.

Assume that a data matrix $\boldsymbol{R}$ (NSA $\times$ NSE) is given, where NSA is the number of samples (objects) and NSE is the number of sensors (variables).

1. Fill a score vector $\boldsymbol{s}$ with suitable starting values (see, e.g., Miyashita et al. [5]).

2. Solve the regression equation

$$\boldsymbol{R} = \boldsymbol{s} \cdot \boldsymbol{l}^{\mathrm{T}} \tag{2}$$

with respect to the loading vector $\boldsymbol{l}$. Normalize $\boldsymbol{l}$.

3. Solve the regression equation

$$\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{l} \cdot \boldsymbol{s}^{\mathrm{T}} \tag{3}$$

with respect to the score vector $\boldsymbol{s}$.

4. If there is no convergence, go to step 2.

5. Set

$$\boldsymbol{E} = \boldsymbol{R} - \boldsymbol{s} \cdot \boldsymbol{l}^{\mathrm{T}} \tag{4}$$

6. If more components are wanted, set

$$\boldsymbol{R} = \boldsymbol{E} \tag{5}$$

and return to step 1.

The NIPALS algorithm for one component may be thought of as a one-dimensional alternating regression where one column of the score and loading matrices are extracted one pair at the time. Such an alternating regression can be done with several or all the components taken together [6]. The algorithm is then as follows.

Assume that a data matrix $\boldsymbol{R}$(NSA × NSE) is given. NMO is the number of components in the model, ideally equal to the number NCO of feasible components in the data.

1. Fill the score matrix $\boldsymbol{S}$(NSA × NMO) with suitable starting values. Karjalainen and Karjalainen [6] suggested simply using random values.

2. Solve the regression equation

$$\boldsymbol{R} = \boldsymbol{S} \cdot \boldsymbol{L}^{\mathrm{T}} \tag{6}$$

for the loading matrix $\boldsymbol{L}$(NSE × NMO). The columns of $\boldsymbol{L}$ may be normalized.

3. Solve the regression equation

$$\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{L} \cdot \boldsymbol{S}^{\mathrm{T}} \tag{7}$$

with respect to the score matrix $\boldsymbol{S}$.

4. If there is no convergence, go to step 2.

The regression equations in the two algorithms above are usually solved using least squares. A comparison of the last score matrix and the one from the previous iteration is often a suitable convergence criterion.

INFLUENCE OF OUTLIERS ON LEAST-SQUARES REGRESSION

It has usually been assumed that the least-squares model gives a good representation of the majority of the data, while one or more outliers are marked by large residuals. This is not always the case. In some situations, outliers will have a disproportionate influence on the model, sometimes exerting such a pull that they end up with lower residuals than some of the points in the majority of the data set. In such situations, estimators with less sensitivity to outliers can be useful alternatives. Several examples were given by Rousseeuw and Leroy [2].

SOME ROBUST REGRESSION METHODS

Given the linear model

$$\boldsymbol{y}(n) = \boldsymbol{X}(n \times p) \cdot \boldsymbol{\theta}(p) \tag{8}$$

where n is the number of data points and p is the number of regression coefficients, one wishes to use a regression estimator which is robust to outliers in both $\boldsymbol{X}$ and $\boldsymbol{y}$. The sensitivity of an estimator to outliers is often measured by the breakdown point, which can be defined as the fraction (or percentage) of data points needed to carry the estimator over all bounds [2]. The least-squares (LS) estimator:

$$\underset{\theta}{\mathit{Minimize}} \sum_{i=1}^{n} r_i^2 \tag{9}$$

($r_i = \hat{y}_i - y_i$ being the difference between the observed and predicted values of y_i,) has a breakdown point of 0%, which means that it can be seriously affected by a single outlier. Several robust estimators having a breakdown point as high as 50% have been proposed. Two of the most common are the least median of squares (LMS), given by

$$\underset{\theta}{\mathit{Minimize}}\ \underset{i}{\mathit{median}}\left(r_i^2\right) \tag{10}$$

and the least trimmed squares (LTS) estimator, defined by

$$\underset{\theta}{\mathit{Minimize}} \sum_{i=1}^{h} \left(r^2\right)_{i:n} \tag{11}$$

where $(r^2)_{1:n} \leqslant \ldots \leqslant (r^2)_{n:n}$ are the first squared, then sorted residuals, and h is usually set to $n/2$. These two estimators are very robust to outliers, but there is a price to be paid as the computational burden is much higher than for least-squares regression, and the algorithms are much less straightforward.

A ROBUST ALTERNATING REGRESSION METHOD

A combination of the alternating regression method and the LMS regression estimator results in the following algorithm.

Run the alternating regression algorithm as given above using LMS regression. Each regression is done by a function called $\langle LMS\_matrix \rangle$ which takes two matrices A and B as parameters, and returns a matrix X. This routine is called in step 2 in the alternating regression as $\langle L^T = LMS\_matrix(R, S) \rangle$, and in step 3 as $\langle S^T = LMS\_matrix(R^T, L) \rangle$.

$\langle LMS\_matrix(B, A) \rangle$ performs the following steps:

A matrix regression equation $B = A \cdot X$ is to be solved.

For every column b_i in B, do

1. $x = LMS\_vector(b_i, A)$; see below.
2. Column i in X is set equal to x.

$\langle LMS\_vector(b, A) \rangle$ performs the following steps:

A vector regression equation $b = A \cdot x$ is to be solved.

1. The dimension of x is set to p = the number of columns in A.
2. Form A_p containing p rows from A, and b_p containing the corresponding p rows from b.
3. Solve $b_p = A_p \cdot x$
4. Calculate M = median $(b_i - (A \cdot x)_i)$ taken over all i.
5. If M is a minimum, return x, or else go to step 2 and use a new subset of lines from A and b.

The current implementation of $\langle LMS\_vector(b, A) \rangle$ does not check all the $\binom{n}{p}$ subsets, as this would make the computational burden impractical or prohibitive. It starts instead with one subset and changes one line in that set following a simple indexing scheme. It is thus not guaranteed that the returned vector x is the optimum solution vector, but there will be a high probability that the final subset $\{A_p, b_p\}$ contains no outliers. The alternating regression phase produces a score matrix S and a loading matrix L. These matrices are not column orthogonal like the ones produced by NIPALS or LS alternating regression. S and L can, however, be multiplied to give an outlier-corrected data matrix R_{SL}. Alternatively, the outliers or suspected outliers may be removed from the data set before further processing.

THE STOP CRITERION PROBLEM

The above algorithm was implemented in the Borland Turbo C++ language, and numerous runs seem to indicate that a simple general stop criterion is hard to find. This problem is due to the fact that the elements in S and L will start to "wobble" towards the end of the iteration process. A comparison of the last and the previous score matrix was tried but, because of the perturbations in S and L, a problem-independent optimum tolerance parameter is also hard to find. Such a tolerance parameter seems to be determined by the signal-to-noise ratio, which is highly dependent on the data set at hand. We also tried to check whether the entity M computed in the function $\langle LMS\_vector(b, A) \rangle$ was less then a tolerance parameter, but this criterion faced the problems stated above.

Rousseeuw [7] suggested the following computation of the LMS scale estimate:

$$\sigma = 1.483 \sqrt{\underset{i=1,\ldots,n}{median}(r_i^2)} \qquad (12)$$

where the constant 1.483 makes σ a consistent estimator of the standard deviation. The LMS scale estimate is robust with a breakdown point of 50%, and if the absolute value of the standardized residual r_i/σ is larger than, e.g., 2.5, data point No. i can be considered to be an outlier.

An alternative approach is to plot the residuals after each iteration and to look for tendencies in the residual structure. If the absolute values of

some of the residuals are towering above the majority, then these residuals indicate outliers.

EXAMPLES

The current implementation of the alternating LMS regression algorithm was tested on a number of different data sets. Randomly selected elements in the data matrix had their values changed to become outliers, and the number of iterations and the time that the algorithm required were recorded. Both the number of iterations and the execution time depend heavily on the stop criterion. All simulations shown in Table 1 and later were made on a Brick PC having a 486DX 33-MHz processor.

It should also be mentioned that successive extraction of components using a robust version of NIPALS (one-component alternating regression) has been tried, so far with unsatisfactory results.

Rows with "Error" in the time column denote runs where the final result was wrong owing to problems with the stop criterion. The criterion used in these runs was as follows: take the mean of the largest least-squares medians from steps 2 and 3 of one iteration, and compare it with the mean obtained in the previous iteration. If the new mean is larger, it is taken to signify that the process has started to wobble around an optimum. For some data sets, however, this criterion results in a premature stop. This may be avoided to some extent by performing a few additional iteration cycles.

Some numerical experiments were also done applying a higher number NMO of components than the actual number of components NCO in the synthetic data set. The algorithm will produce a score matrix S with some extra columns that

TABLE 1

Run times for some selected examples.

NSA	NSE	NMO	Outliers	Iterations	Time (s)
10	10	1	0	4	2.5
10	10	1	1	5	3.1
10	10	1	3	5	3.1
10	10	1	5	5	3.2
10	10	2	0	4	5.8
10	10	2	1	5	7.4
10	10	2	3	4	5.5
10	10	2	5	4	5.5
10	10	3	0	7	17.5
10	10	3	1	5	12.2
10	10	3	3	5	12.0
10	10	3	5	4	Error
30	30	1	0	4	53.6
30	30	1	1	5	68.8
30	30	1	3	6	84.1
30	30	1	5	5	68.8
30	30	2	0	6	190.7
30	30	2	1	8	259.8
30	30	2	3	7	225.2
30	30	2	5	6	192.5
30	30	3	0	6	329.1
30	30	3	1	7	386.8
30	30	3	3	11	625.9
30	30	3	5	6	Error
60	60	3	5	9	2918.0

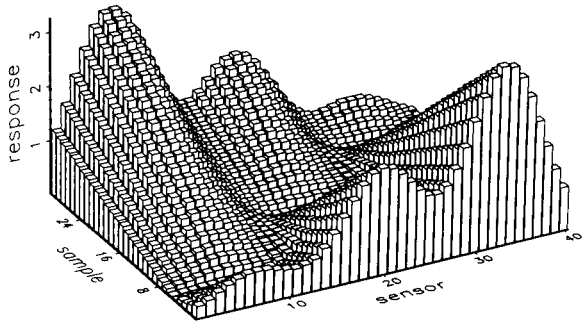


Fig. 1. Synthetic data set having three components and normally distributed noise having a standard deviation of 0.01 response units. The response is shown as a function of sensor and sample indices.

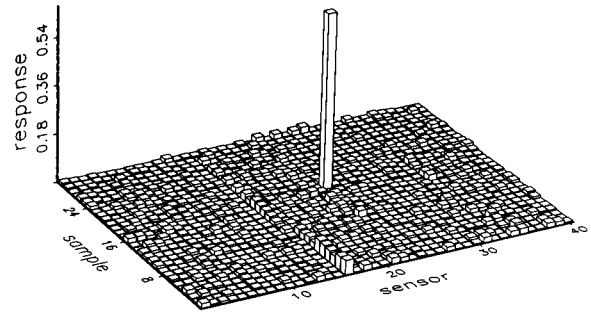


Fig. 3. Plot of the absolute values of the residuals obtained from running ten alternating robust regression iterations on the data set in Fig. 2. The outlier towers high above the remainder of the residuals.

also model noise, and likewise with the loading matrix $\boldsymbol{L}$. Nevertheless, $\boldsymbol{R}_{SL} = \boldsymbol{S} \cdot \boldsymbol{L}$ still produces an outlier-free matrix. A set of orthogonal components may then be generated fairly quickly by principal component or singular value decomposition. If NMO is set too small, relevant information contained in the original data matrix $\boldsymbol{R}$ will be lost.

Finally, a few instances of outliers in a synthetic three-component data set with normally distributed noise with standard deviation 0.01 response units added (Fig. 1) will be shown.

First a single data element is chosen at random and made an outlier (Fig. 2). The absolute values of the residuals after a three-component alternating regression is shown (Fig. 3) where the outlier is clearly visible. A back multiplication to $\boldsymbol{R}_{SL}$ yields a response matrix indistinguishable from Fig. 1.

Second, eight randomly chosen data elements are chosen and made outliers (Fig. 4). Once again

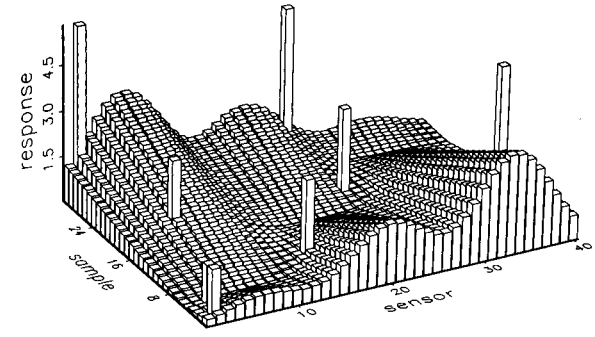


Fig. 4. Data set in Fig. 1 with different offsets applied to eight randomly chosen elements.

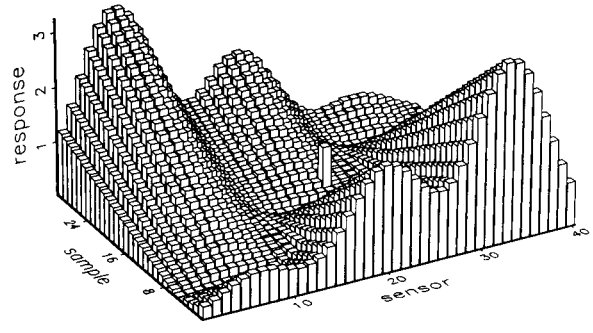


Fig. 2. Data set in Fig. 1 with the element (23,18) having a 0.64 response unit offset to become an outlier.

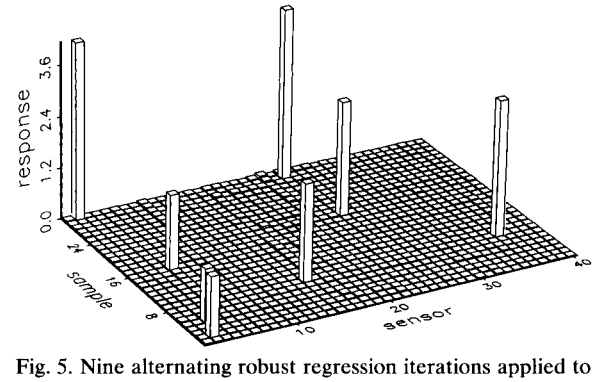


Fig. 5. Nine alternating robust regression iterations applied to the data set in Fig. 4. The outliers are clearly visible on this absolute value residual plot.

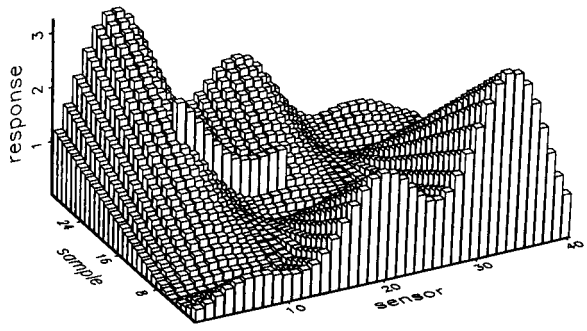


Fig. 6. Data set in Fig. 1 with elements (8,19) to (19,19) having a 0.6 response unit offset.

the outliers are clearly seen in the four-component residual plot (Fig. 5) and $\boldsymbol{R}_{SL}$ is faithfully reproduced by the multiplication of scores and loadings.

Finally several successive elements in a row of the data matrix are offset (Fig. 6). Even in this case the outliers are clearly seen in a three-component residual plot. However, in this case there are several other residual values of comparable magnitude forming a ridge on a single row of the residual matrix (Fig. 7).

It seems from this and several other tests, among others one where a full row in the data matrix is offset, that structured rather than isolated outliers will force a modification of the overall model even with this robust regression algorithm. An inspection of the residual plots will even in more questionable cases yield important information about the data structure. The problem of structured outliers is the subject of further study.

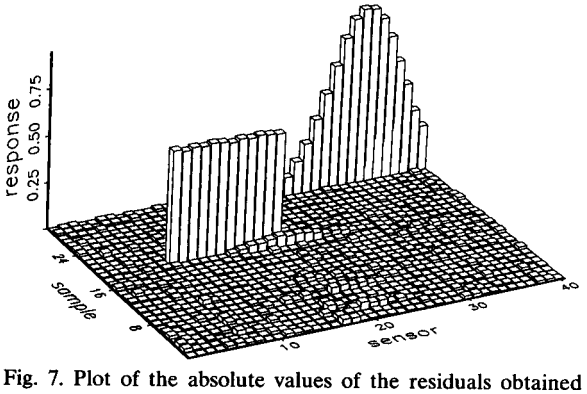


Fig. 7. Plot of the absolute values of the residuals obtained from running six alternating robust regression iterations on the data set in Fig. 6. The offset ridge is readily seen, but another ridge indicates problems with the algorithm.

DISCUSSION

The algorithm given seems to identify outliers fairly well, in particular with visual inspection of the residuals, and can be to remove the outliers from the data set. Isolated outliers are easily recognized, although it will always be a matter of judgement whether a data element contains a unusually large noise component or should be considered an outlier.

The major problem that is the subject of further investigation is an improved regression algorithm, with respect to both the speed of computation and structured outliers. A solution could be a better algorithm for computing the LMS alternating regression, or the application of a different estimator than the LMS. Rousseeuw and Leroy [2] described various robust estimators worth examining.

Whichever regression algorithm is implemented, it will be necessary to find stable criteria both for terminating the iterations and recognizing outliers. For unsupervised use the termination criterion should preferably be as stable as the comparison of vectors or matrices during successive iteration is for LS regression.

REFERENCES

1 H. Martens and T. Næs, Multivariate Calibration, Wiley, New York, 1989.
2 P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
3 H. Wold, in F. David (Ed.), Research Papers is Statistics, Wiley, New York, 1966, p. 411.
4 S. Wold, C. Albano, W.J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson and M. Sjöström, in H. Martens and H. Russwurm (Eds.), Food Research and Data Analysis, Applied Science, London, 1983, p. 147.
5 Y. Miyashita, T. Itozawa, H. Katsumi and S.I. Sasaki, J. Chemometr., 4 (1990) 97.
6 E.J. Karjalainen and U.P. Karjalainen, Clin. Chem. Res. Found. Lib., 2 (1987) 1.
7 P.J. Rousseeuw, J. Chemometr., 5 (1991) 1.

Analytica Chimica Acta, 277 (1993) 495–501
Elsevier Science Publishers B.V., Amsterdam

An approach to interval estimation in partial least squares regression

A. Phatak, P.M. Reilly and A. Penlidis
Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1 (Canada)

(Received 3rd September 1992)

Abstract

Although partial least squares regression (PLS) is widely used in chemometrics for quantitative spectral analysis, little is known about the distribution of the prediction error from calibration models based on PLS. As a result, we must rely on computationally intensive procedures like bootstrapping to produce confidence intervals for predictions, or, in many cases, we must do with no interval estimates at all, only point estimates. In this paper we present an approach, based on the linearization of the PLS estimator, that allows us to construct approximate confidence intervals for predictions from PLS.

Keywords: Interval estimation; Partial least squares regression

Partial least squares regression (PLS) [1–3] is a biased regression method that is often used in quantitative spectral analysis. Because of the highly non-linear form of the PLS estimator, however, it is very difficult to derive the exact distribution of the prediction error from calibration models based on PLS. Thus, we must rely on computationally intensive procedures like bootstrapping to produce confidence intervals for predictions from PLS, or, in many cases, we must do with no interval estimates at all, only point estimates. In this paper, we show that it is possible to linearize the PLS estimator and thereby obtain its approximate covariance matrix. As a result, we can produce approximate confidence intervals for predictions from PLS.

The paper consists of four parts: first, a short discussion of PLS; second, an equally brief description of how the linearization of the PLS estimator was carried out; third, a discussion of the results of Monte Carlo simulations whose goal was to determine how close the approximate covariance matrix was to the true covariance matrix; and finally, an illustration, using spectroscopic data published by Fearn [4], of how confidence intervals for predictions from PLS can be constructed.

Correspondence to: A. Phatak, Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1 (Canada).

PARTIAL LEAST SQUARES REGRESSION

In this section, we briefly describe PLS and its relationship to ordinary least squares (OLS). Many good expositions of PLS have appeared in the literature [2,5,6], and we refer the interested reader to these articles for a more detailed discussion of the mechanics of the PLS algorithms.

To begin with, we consider the standard regression model defined by the equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of observations on the dependent variable, $\mathbf{X}$ is an $(n \times p)$ matrix whose (i,j)th element is the value of the jth independent variable for the ith observation, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters, and $\boldsymbol{\epsilon}$ is an $(n \times 1)$ vector of errors identically and independently distributed with mean zero and variance σ^2. For the sake of convenience, we assume that both $\mathbf{y}$ and $\mathbf{X}$ have been mean-centered and that $n > p$. In the calibration of a near-infrared (NIR) reflectance instrument, for example, each row of $\mathbf{X}$ would consist of reflectances at p wavelengths in some specified range of a sample containing a measured amount, y_i, of the compound of interest. The objective of the statistical analysis, then, is to estimate the elements of $\boldsymbol{\beta}$ so that we can predict the composition of a new sample from its NIR spectrum.

The ordinary least squares estimate of $\boldsymbol{\beta}$ in Eqn. 1 is given by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2}$$

Because $\hat{\boldsymbol{\beta}}_{OLS}$ is a linear function of $\mathbf{y}$, its covariance matrix can be calculated as

$$\begin{aligned} \mathrm{var}(\hat{\boldsymbol{\beta}}_{OLS}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathrm{var}(\mathbf{y})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \end{aligned} \tag{3}$$

where var($\mathbf{y}$), the covariance matrix of $\mathbf{y}$, is equal to $\mathbf{I}\sigma^2$. Having determined var($\hat{\boldsymbol{\beta}}_{OLS}$), we can then go on to calculate the variance of a prediction, $\hat{y}$, from a new observation, say $\tilde{\mathbf{x}}(p \times 1)$, as

$$\mathrm{var}(\hat{y}) = \left[\tilde{\mathbf{x}}'(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{x}}\right]\sigma^2 \tag{4}$$

We can see from Eqns. 3 and 4, therefore, the link between the variance of the *estimate* and the variance of a *prediction*.

When the columns of $\mathbf{X}$ are highly collinear, as they will be when $\mathbf{X}$ consists of digitized spectra, the variances of the individual coefficient estimates will be very large and the predictor based on these estimates may give very poor predictions [7]. Biased regression methods such as ridge regression [8], principal component regression (PCR) [9], and partial least squares overcome many of the deficiencies of least squares in these circumstances and are therefore often used in its place.

In PLS, the estimator of $\boldsymbol{\beta}$ in Eqn. 1 can be written as

$$\hat{\boldsymbol{\beta}}^m_{PLS} = \mathbf{V}_m(\mathbf{V}_m'\mathbf{X}'\mathbf{X}\mathbf{V}_m)^{-1}\mathbf{V}_m'\mathbf{X}'\mathbf{y} \tag{5}$$

where m denotes the number of PLS "dimensions" retained in the calibration model. It is usually determined by cross-validation [1]. In Eqn. 5, the matrix $\mathbf{V}_m$ is *any* $(p \times m)$ matrix whose columns form a basis [2] for the space spanned by

$$\left[\mathbf{X}'\mathbf{y} \,|\, (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y} \,|\, \ldots \,|\, (\mathbf{X}'\mathbf{X})^{(m-1)}\mathbf{X}'\mathbf{y}\right]$$

In practice, the columns of $\mathbf{V}_m$ are usually the orthogonal, unit-norm PLS weight vectors (denoted by $\mathbf{w}_i$ in the chemometric literature) obtained from the PLS algorithm of Wold et al. [10]. Thus, unlike the least squares estimator, which is a non-stochastic transformation of the random variable $\mathbf{y}$, we can see that the PLS estimator is a highly non-linear function of $\mathbf{y}$. As a result, we cannot determine its covariance matrix as easily as we did in Eqn. 3 for the OLS estimator. We show in the next section, however, that it is possible to *linearize* the PLS estimator, $\hat{\boldsymbol{\beta}}^m_{PLS}$, and by doing so, come up with an approximate covariance matrix.

LINEARIZING THE PLS ESTIMATOR

In this section, we outline how the linearized form of the PLS estimator was derived and then state the result. The details of the proof will appear in a forthcoming publication [11]. We should point out here that the general method of linearizing a non-linear function of a random variable to estimate its variance is not new to analytical chemists (see, for example, Skoog and West [12, pp. 76–80]).

Linearizing the PLS estimator in Eqn. 5 amounts to writing it out as a Taylor series expansion and then truncating it after a single term, i.e.,

$$\hat{\boldsymbol{\beta}}^m_{PLS} \approx \left(\hat{\boldsymbol{\beta}}^m_{PLS}\right)_o + \mathbf{J}_o(\mathbf{y} - \mathbf{y}_o) \tag{6}$$

where the subscript $(_o)$ denotes that the linearization has been carried out *around the data*, $\mathbf{y}_o$ and $\mathbf{X}_o$. Thus the Jacobian matrix, $\mathbf{J}$, which is the

matrix of derivatives of each element of $\hat{\beta}^m_{PLS}$ with respect to the elements of **y**, is also evaluated at the data. Then, taking the variance of both sides of Eqn. 6, the approximate covariance matrix of $\hat{\beta}^m_{PLS}$ given the data is simply

$$\mathrm{var}(\hat{\beta}^m_{PLS}) \approx \mathbf{J}_o\mathbf{J}_o'\sigma^2 \quad (7)$$

At first glance, coming up with a general expression for **J** seems an intimidating task, and indeed it is, but if we adopt the approach of Magnus and Neudecker [13] to matrix differential calculus, deriving the Jacobian matrix is quite straightforward. Below, we give the general expression for **J**. To simplify its presentation, however, we first introduce some notation. Let $\mathbf{s} = \mathbf{X}'\mathbf{y}$, $\mathbf{S} = \mathbf{X}'\mathbf{X}$, $\mathbf{R}_m = (\mathbf{V}_m'\mathbf{S}\mathbf{V}_m)^{-1}$, and $\mathbf{H}_m = \mathbf{V}_m\mathbf{R}_m\mathbf{V}_m'$. Furthermore, define $\mathbf{I}_p$ to be the $(p \times p)$ identity matrix and $\mathbf{U}_m$ to be the $(n \times mp)$ matrix

$$\mathbf{U}_m = [\mathbf{X}\,|\,\mathbf{X}\mathbf{S}\,|\,\ldots\,|\,\mathbf{X}\mathbf{S}^{(m-1)}] \quad (8)$$

Finally, let **M** be an $(m \times m)$ non-singular matrix defined by the equation

$$\left[\mathbf{X}'\mathbf{y}\,|\,(\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y}\,|\,\ldots\,|\,(\mathbf{X}'\mathbf{X})^{(m-1)}\mathbf{X}'\mathbf{y}\right] = \mathbf{V}_m\mathbf{M} \quad (9)$$

Then, the Jacobian matrix, **J**, can be written as

$$\mathbf{J} = \left\{\left[\mathbf{s}'\mathbf{V}_m\mathbf{R}_m \otimes (\mathbf{I}_p - \mathbf{H}_m\mathbf{S})\right] + \left[\mathbf{V}_m\mathbf{R}_m \otimes \mathbf{s}'(\mathbf{I}_p - \mathbf{H}_m\mathbf{S})\right]\right\}(\mathbf{M}^{-1'} \otimes \mathbf{I}_p)\mathbf{U}_m' + \mathbf{H}_m\mathbf{X}' \quad (10)$$

where the symbol $\otimes$ denotes the Kronecker product. The Jacobian matrix in Eqn. 10 consists of two terms. The second term, $\mathbf{H}_m\mathbf{X}'$, can be thought of as a first-order approximation to the Jacobian matrix itself. We can derive it very easily if we assume that the space spanned by the columns of $\mathbf{V}_m$ changes little as **y** changes. In other words, $\mathbf{H}_m\mathbf{X}'$ is the derivative of $\hat{\beta}^m_{PLS}$ with respect to **y** treating $\mathbf{V}_m$ as fixed (non-stochastic). The first term then is a refinement to this first-order approximation. We note that it depends on the matrix **M**, defined in Eqn. 9. Thus, Eqn. 10 is completely general inasmuch as we can choose any $\mathbf{V}_m$ we like to calculate the Jacobian matrix, as long as it satisfies Eqn. 9. In practice, of course, **J** is evaluated using the data, $\mathbf{y}_o$ and $\mathbf{X}_o$.

How good is the approximation $\mathbf{J}_o\mathbf{J}_o'\sigma^2$ to the true covariance matrix of the PLS estimator? In the next section, we summarize the results of a preliminary Monte Carlo study whose goal was to assess the usefulness of the approximation.

MONTE CARLO STUDY

The Monte Carlo study we carried out was based on a similar one by Gunst and Mason [14], who compared the performance of several biased regression methods. Ours, however, is a preliminary study and is restricted to PLS.

The simulations were based on the linear model described by Eqn. 1. We generated a (30×10) **X**-matrix consisting of uniformly distributed numbers in the range (0,1) and then mean-centered and variance-scaled each column. The last five columns of **X** were generated in such a way that they were near linear combinations of the first five; as a result, the eigenvalues of $\mathbf{X}'\mathbf{X}$ were

$$(l_1, l_2, \ldots, l_{10}) = (3.13, 2.82, 2.01, 1.14, 0.90, 0.0024, 0.0014, 0.0011, 0.00035, 0.00032)$$

In the two sets of simulations described below, different values of σ^2 and different vectors of coefficients, β, were used. In the first set, which we denote by S1, $\sigma^2 = 0.01$ and $\beta = 0.5(z_1 + z_2 + z_9 + z_{10})$ where the z_i, $i = 1, 2, \ldots, 10$ are the

TABLE 1

Diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$

i									
1	2	3	4	5	6	7	8	9	10
1270.2	715.8	182.7	1201.9	622.6	1164.7	538.6	305.1	624.5	1376.5

TABLE 2
Summary of results from S1 ($\sigma^2 = 0.01$)

i	β	$\sigma(\hat{\beta}_{OLS})$	PLS ($m = 3$)		
			$E(\hat{\beta}_{PLS})$	$\sigma(\hat{\beta}_{PLS})$ (true)	$\sigma(\hat{\beta}_{PLS})$ (lin.)
1	0.59	3.56	0.19	0.052	0.053
2	−0.28	2.68	−0.04	0.075	0.076
3	−0.18	1.35	−0.14	0.084	0.085
4	0.49	3.47	0.32	0.053	0.055
5	−0.26	2.50	−0.09	0.066	0.067
6	0.06	3.41	0.04	0.037	0.038
7	−0.38	2.32	−0.38	0.042	0.044
8	−0.19	1.75	−0.22	0.048	0.051
9	−0.17	2.50	0.05	0.046	0.047
10	0.06	3.71	−0.36	0.042	0.042

unit-norm eigenvectors of $\mathbf{X}'\mathbf{X}$. In the second set, S2, $\sigma^2 = 0.1$ while $\boldsymbol{\beta}$ was an arbitrary linear combination of *all* the z_i and, as in the first set, it was scaled so that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. Having established $\mathbf{X}$ and $\boldsymbol{\beta}$, we could then generate $\mathbf{y}$ according to Eqn. 1 and $\mathbf{J}_o$ from Eqn. 10. Over the course of 10 000 trials, the average squared coefficient of correlation in S1 was 0.95, while in S2 it was 0.71. Before discussing the results of the simulations, it is worthwhile to examine the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ (Table 1); when multiplied by σ^2, these will give the variances of the coefficients estimated by least squares. Not surprisingly some of the diagonal elements are extremely large, which is a direct consequence of the five near-zero eigenvalues of $\mathbf{X}'\mathbf{X}$.

S1 ($\sigma^2 = 0.01$, $r^2_{avg} = 0.95$)

Some of the results from S1 are summarized in Table 2. Cross-validation throughout the trials indicated that only two or three dimensions needed to be included in the model and hence we report results for $m = 3$ only. The first column shows the true $\boldsymbol{\beta}$, constructed as described above. The third column shows the mean value of 10 000 trials of the PLS estimator. After such a large number of trials, this can be considered the expected value of $\hat{\boldsymbol{\beta}}^3_{PLS}$. We note that it is different from the true $\boldsymbol{\beta}$: therein lies the meaning of the term "bias". The fourth column gives the standard deviations of the elements of the PLS estimator. They are almost two orders of magnitude smaller than their counterparts in OLS, which are shown in the second column. This is not a surprising result; like principal component regression, PLS deletes directions in the $\mathbf{X}$-space associated with smaller eigenvalues. However, because PLS variates are chosen by taking $\mathbf{y}$ into account the degree of variance reduction is not as large as in PCR for an equal number of dimensions extracted. The fifth column also shows the standard deviations of the elements of $\hat{\boldsymbol{\beta}}^3_{PLS}$, but the figures here represent the average *over all trials* of the diagonal elements of $\mathbf{J}_o\mathbf{J}_o'\sigma^2$. Agreement with column 4 is excellent as we would expect with such a small variance. In other words, for $\sigma^2 = 0.01$, the linearization of the PLS estimator in Eqn. 6 is a good approximation. How close are the diagonal elements of $\mathbf{J}_o\mathbf{J}_o'\sigma^2$ from any *single* trial to the averages in the last column of Table 2? In the 10 000 trials that were carried out, the largest difference in the standard deviations was only ± 0.005 units. As we will see in the next section, however, the distribution of the variance or standard deviation of any individual coefficient will not always be this narrow.

S2 ($\sigma^2 = 0.1$, $r^2_{avg} = 0.71$)

In the second set of simulations, we increased the variance ten-fold, to a value of $\sigma^2 = 0.1$. In addition, the orientation of $\boldsymbol{\beta}$ was altered so that is was an arbitrary linear combination of all the eigenvectors of $\mathbf{X}'\mathbf{X}$. Nevertheless, cross-validations throughout the trials indicated again that only two or three dimensions were required.

The results of S2 are summarized in Table 3. Many of the observations we made about the results of S1 can also be made here, for example, the biasedness of the PLS estimate of the vector of coefficients and the small size of the standard deviations of its elements relative to those of the least squares estimate. As in S1, we generated for each trial the approximate covariance matrix $\mathbf{J}_o\mathbf{J}_o'\sigma^2$ and then extracted its diagonal elements, which represent the variances of the individual coefficients. The figures in column 6 of Table 3 are the medians over all trials of the standard deviations of the coefficients of $\hat{\boldsymbol{\beta}}^3_{PLS}$. We report the median instead of the mean because the

TABLE 3

Summary of results from S2 ($\sigma^2 = 0.1$)

i	β	$\sigma(\hat{\beta}_{OLS})$	PLS ($m = 3$)		
			$E(\hat{\beta}_{PLS})$	$\sigma(\hat{\beta}_{PLS})$ (true)	$\sigma(\hat{\beta}_{PLS})$ (lin.)
1	0.36	11.27	0.11	0.17	0.16
2	0.06	8.46	0.05	0.24	0.24
3	0.23	4.27	0.34	0.27	0.27
4	0.85	10.96	0.43	0.17	0.17
5	0.13	7.89	0.08	0.21	0.21
6	0.16	10.79	0.17	0.12	0.12
7	−0.01	7.34	−0.35	0.13	0.13
8	0.14	5.52	−0.15	0.15	0.15
9	−0.07	7.90	0.08	0.15	0.15
10	−0.16	11.73	−0.14	0.13	0.13

distributions of the standard deviations of these coefficients are slightly skewed (see, for example, Fig. 1). Again, agreement is excellent. We should point out, however, that of the 10000 trials a few yielded standard deviations that were very large. For example, in Fig. 1 we show the distribution of the standard deviation of the first element of $\hat{\beta}^3_{PLS}$ obtained from the approximate covariance matrices generated during the trials. Almost all (9980) of the observations are between 0.14 and 0.3; however, 19 are between 0.3 and 1.0, while a single, "pathological" value of 12.9 was also generated. At the present time, we cannot explain why such large values occur. However, a plausible explanation is that in those instances which yield large variances, we are using too many dimensions. Recall that the Monte Carlo trials were carried out for a fixed number of dimensions ($m = 3$). It is possible that for the vectors of observations, $\mathbf{y}$, corresponding to extreme values a model with $m = 1$ or 2 might have been more appropriate. We base this tentative explanation on the observation that as we increase the number of dimensions well beyond what might be chosen by cross-validation, that is, as we overfit the model, the distribution of the variances or standard deviations of the individual coefficients becomes broader and more skewed.

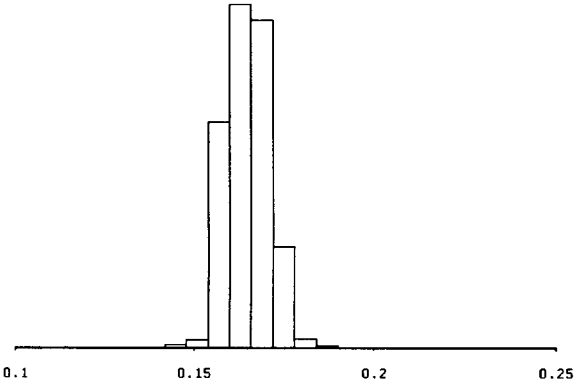


Fig. 1. Distribution of estimated standard deviations of $(\hat{\beta}^3_{PLS})_1$. Histogram consists of 9980 observations.

For each of the 10000 trials, we also calculated the residual sum of squares, defined as $RSS = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$. For a PLS model using three dimensions, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}^3_{PLS}$, and for OLS, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$. The average RSS from OLS was 1.889, which when divided by the appropriate number of degrees of freedom ($30 - 10 - 1 = 19$) yields, as expected, a number close to the true variance, 0.1. Estimating σ^2 from the PLS residual sum of squares is a much more difficult problem and it is further complicated by the biasedness of the estimator. The average RSS for PLS with $m = 3$ was 2.47. If we divide this by ($n - m - 1 = 26$), we slightly *underestimate* the variance. On the other hand, if we divide by 19, the number of degrees of freedom in least squares, we *overestimate* the variance. Dividing by ($n - m - 1$) is appealing, if for no other reason than by doing so we err on the side of caution. If, however, m, the number of dimensions retained in the PLS model, is much less than p, the number of x-variables, we might severely overestimate σ^2. On the other hand, if the bias is not too large, then dividing by ($n - p - 1$) may, in the long run, only slightly underestimate the variance. Gunst and Mason [15, pp. 323–329], in a discussion about analysis of variance and inference in principal component regression, suggest yet another alternative: simply using the OLS estimate of the variance.

The results of the simulations indicate that the covariance matrix of the PLS estimator based on the Jacobian matrix does provide a useful approximation to the true covariance matrix. Because the Jacobian matrix is evaluated using the data from a single experiment, however, how well any single calculation of $\mathbf{J}_o\mathbf{J}_o'\sigma^2$ approximates the

TABLE 4

Comparison of coefficient estimates from OLS and PLS (data from Fearn [4])

i	OLS		PLS ($m = 4$)	
	$\hat{\beta}$	S.D.	$\hat{\beta}$	S.D.
1	0.311	3.04	0.003	0.022
2	0.015	2.69	0.128	0.013
3	0.011	3.45	0.133	0.013
4	−0.441	2.18	−0.209	0.015
5	0.021	0.22	0.007	0.006
6	0.138	1.82	−0.052	0.040

true covariance matrix depends on fitting just the right number of dimensions.

ILLUSTRATION USING SPECTROSCOPIC DATA

In this section, we briefly illustrate how Eqn. 7 can be used to construct approximate confidence intervals for predictions from PLS. We use the spectroscopic data published by Fearn [4]. The dataset consists of (a) measurements of reflectance of NIR radiation at six different wavelengths by samples of wheat and (b) the protein content of these samples. The calibration set contains 24 samples while the test set consists of 26 samples. Naes and Martens [7], using only the first 12 samples of the calibration data, showed that a PLS model of four dimensions yielded the best predictions. We also based our analysis on the same reduced calibration set and on a PLS model of the same dimensionality. For the PLS analysis, the data were mean-centered only.

Table 4 shows the estimates of the vector of coefficients, $\boldsymbol{\beta}$, from both PLS and OLS as well as their corresponding standard deviations. The standard deviations of the individual coefficients were calculated as the square roots of the diagonal elements of the matrices $(\mathbf{X}'\mathbf{X})^{-1}s^2$ and $\mathbf{J}_o\mathbf{J}_o's^2$, respectively, for OLS and PLS. To estimate σ^2 in PLS, we divided the residual sum of squares by $(n - m - 1)$. This gave a result of $s^2 = 0.07$. The corresponding estimate in OLS was 0.05. As Table 4 shows, not only is the PLS estimate of $\boldsymbol{\beta}$ very different from the OLS one, but it is also much more precise. Furthermore, Martens and Naes [7] demonstrated that *predictions* using PLS were also more *accurate*. For the entire test set, the root-mean-squared error of prediction from PLS (0.3) was much smaller than the corresponding value from OLS (0.8). Thus, although PLS is a biased method, the bias in this instance is probably quite small.

In the first two columns of Table 5 we show the standard deviations of the predictions for the first ten samples of the test set. These were calculated using Eqn. 4, with $\mathbf{J}_o\mathbf{J}_o's^2$ in place of $(\mathbf{X}'\mathbf{X})^{-1}s^2$ for PLS. In general, the standard deviations from PLS are smaller than those from OLS. As a result, the $\pm 2 \times$ S.D. limits are narrower than the limits from least squares, (columns 3, 5, 6, 8), yet most of the true values of the

TABLE 5

Comparison of $\pm 2 \times$ S.D. limits from OLS and PLS (data from Fearn [4])

S.D.		PLS			OLS		
PLS	OLS	$-2 \times$ S.D.	y	$+2 \times$ S.D.	$-2 \times$ S.D.	y	$+2 \times$ S.D.
1.24	1.79	6.26	8.96	11.20	8.20	8.96	15.37
0.43	0.57	7.23	7.90	8.95	7.85	7.90	10.12
0.21	0.20	9.33	9.27	10.15	9.32	9.27	10.13
0.23	0.25	11.52	11.77	12.45	11.32	11.77	12.32
0.12	0.17	10.02	9.70	10.49	9.63	9.70	10.32
0.16	0.26	10.81	10.46	11.46	10.75	10.46	11.78
0.13	0.28	10.16	10.17	10.67	9.32	10.17	10.46
0.25	0.26	10.76	11.10	11.77	10.57	11.10	11.59
0.32	0.40	11.20	12.03	12.46	10.69	12.03	12.27
0.69	0.65	7.81	9.43	10.59	8.20	9.43	10.79

protein content for the test set (columns 4 and 7) still lie within those limits.

Conclusion

In this paper, we have outlined an approach that allows us to construct approximate confidence limits for predictions from PLS, and we showed how it could be applied to spectroscopic data. Our approach is based on linearizing the PLS estimator to determine its covariance matrix. The results of a limited number of simulations show that the covariance matrix derived in this way is good approximation to the true covariance matrix; they also indicate that how good it is may depend on fitting just the right number of dimensions. Although the question of how best to estimate the measurement error, σ^2, from PLS must be resolved, we believe that this approach has immediate practical value in providing approximate confidence limits for predictions in a simple, straightforward manner.

We would like to thank Dr. C. Heckler and Dr. R. Swanson of Eastman Kodak, Inc., Rochester, NY for their valuable comments on the general approach outlined in this paper.

REFERENCES

1 H. Martens and T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
2 I. Helland. Commun. Stat. Simul. Comput., 17 (1988) 581.
3 A. Phatak, P.M. Reilly and A. Penlidis, Commun. Stat. Theory Meth., 21 (1992) 1517.
4 T. Fearn, Appl. Stat., 32 (1983) 73.
5 A. Höskuldsson, J. Chemometr., 2 (1988) 211.
6 R. Manne, Chemom. Intell. Lab. Syst., 2 (1987) 187.
7 T. Naes and H. Martens, Commun. Stat. Simul. Comput., 14 (1985) 545.
8 A.E. Hoerl and R.W. Kennard, Technometrics, 12 (1970) 55.
9 I.T. Jolliffe, Principal Component Analysis, Springer Verlag, New York, 1986.
10 S. Wold, H. Martens and H. Wold, in A. Ruhe and B. Kågström (Eds.), Lecture Notes in Mathematics: Matrix Pencils, Springer Verlag, Heidelberg, 1983, p. 286.
11 A. Phatak, P.M. Reilly and A. Penlidis, in preparation.
12 D.A. Skoog and D.M. West, Fundamentals of Analytical Chemistry, Saunders, Philadelphia PA, 4th edn., 1982.
13 J.R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley, New York, 1988.
14 R.F. Gunst and R.L. Mason, J. Am. Stat. Assoc., 72 (1977) 616.
15 R.F. Gunst and R.L. Mason, Regression Analysis and its Applications, Marcel Dekker, New York, 1980.

Analytica Chimica Acta, 277 (1993) 503–510
Elsevier Science Publishers B.V., Amsterdam

Bayesian confidence intervals for the product of three normal means

Malwane M.A. Ananda and Ashok K. Singh
Department of Mathematical Sciences, University of Nevada, Las Vegas, NV 89154 (USA)

G.T. Flatman
U.S. Environmental Protection Agency, Las Vegas, NV 89154 (USA)

(Received 3rd September 1992)

Abstract

The problem of estimating the product of three normal means of three independent normal distributions is considered. These types of estimation problems arise in many environmental applications, such as exposure assessment and risk modeling. Classical confidence interval estimates are available in the literature; we consider the problem from the Bayesian approach using two different proper prior distributions and one non-informative prior distribution. Assuming the quadratic loss, Bayesian estimates and Bayesian confidence intervals are given. Numerical integration or simulation are necessary to evaluate such confidence intervals. Computer programs written in Fortran are given to calculate these confidence intervals. Examples are provided.

Keywords: Bayesian estimation; Credible regions; Confidence intervals; Risk/exposure modeling

Suppose X_1, X_2 and X_3 are three independent, normally distributed random variables with means θ_i and variances σ_i^2, $i = 1, 2, 3$. We are concerned with inference, mainly estimation and confidence statements about the product of their means $\psi = \theta_1\theta_2\theta_3$. These type of estimation problems arise in many environmental applications, such as exposure assessment and risk modeling, where measurements are assumed to have different independent normal populations with unknown means and variances.

Suppose a random sample $x_{i1}, x_{i2}, \ldots, x_{in_i}$ of size n_i $(i = 1, 2, 3)$ is available from each population and let $(\bar{x}_i, s_i)$ be their sample mean and sample standard deviation. Then $\{\bar{x}_1, \bar{x}_2, \bar{x}_3, s_1, s_2, s_3\}$ makes a complete sufficient statistic [1] for $\{\theta_1, \theta_2, \theta_3, \sigma_1, \sigma_2, \sigma_3\}$. Although some risk assessors [2] have proposed that the confidence interval of the product of three means can be calculated as the product of three individual confidence intervals, Yfantis and Flatman [3] showed that this could lead to misleading answers in many cases. Furthermore, assuming $n_1 = n_2 = n_3 (= n)$ they proposed a confidence interval based on the estimator $\hat{\psi}_0 = \sum_{i=1}^{n} x_{1i}x_{2i}x_{3i}/n$ which can be computed by simulation.

Maximum likelihood estimator of ψ is $\hat{\psi}_1 = \bar{x}_1\bar{x}_2\bar{x}_3$. Notice that both estimators $\hat{\psi}_0$ and $\hat{\psi}_1$ are unbiased estimators of ψ and the mean square error of the estimator $\hat{\psi}_0$ is larger than the mean square error of the estimator $\hat{\psi}_1$ (for the proof,

Correspondence to: A.K. Singh, Department of Mathematical Sciences, University of Nevada, Las Vegas, NV 89154 (USA).

see Ref. 3). Hence the confidence intervals based on $\hat{\psi}_1$ must be more efficient (i.e., shorter) than the confidence intervals based on the estimator $\hat{\psi}_0$. Furthermore equal sample sizes are not an assumption for $\bar{x}_1\bar{x}_2\bar{x}_3$. This is the case in many practical situations, due to missing observations or due to the nature of the problem. For these situations, in order to use the first estimator $\hat{\psi}_0$ data with missing entries must be eliminated. Therefore using $\hat{\psi}_1 = \bar{x}_1\bar{x}_2\bar{x}_3$ over $\hat{\psi}_0$ has clear advantages.

In this paper, we show how to incorporate prior knowledge about the unknown parameters in the inference process. We consider several types of prior families and give their Bayesian estimators and Bayesian credible regions (Bayesian version of confidence intervals) along with a simulation program (written in Fortran using IMSL libraries) to implement the proposed method. The Bayesian tools are efficient whenever the prior knowledge about these unknown parameters are available, as is the case in many practical situations.

Also, we consider generalized Bayesian estimators and confidence intervals for ψ based on the non-informative prior (assuming no prior knowledge about the unknown parameters) and give a simulation program to implement the method. Actually these generalized Bayesian estimators, with respect to the non-informative prior, coincide with the previous unbiased estimator $\hat{\psi}_1 = \bar{x}_1\bar{x}_2\bar{x}_3$. Numerical examples are presented for each case.

BAYESIAN ESTIMATORS AND CREDIBLE REGIONS

Method 1

First consider the case with no prior information, i.e., using a non-informative prior. Let $(\theta_1,\theta_2,\theta_3,\sigma_1,\sigma_2,\sigma_3)$ has the joint improper non-informative prior density

$$\pi(\theta_1,\theta_2,\theta_3,\sigma_1,\sigma_2,\sigma_3) = 1/(\sigma_1\sigma_2\sigma_3);\ 0 < \sigma_i < \infty,\ -\infty < \theta_i < \infty,\ i = 1, 2, 3 \quad (1)$$

Notice here that this non-informative prior is the product of three non-informative priors $\pi_i(\theta_i,\sigma_i) = 1/(\sigma_i)$, which is known as the right invariant Haar density for the one population situation.

Theorem 1

For the prior given (Eqn. 1), the marginal posterior distribution of θ_i given $(\bar{x}_i, s_i)$

$$\theta_i/(\bar{x}_i,s_i) \sim T\left(n_i - 1,\ \bar{x}_i,\ s_i^2/n_i\right);\ i = 1, 2, 3 \quad (2)$$

and the joint posterior distribution of $(\theta_1,\theta_2,\theta_3)$ given $\{(\bar{x}_i,s_i);\ i = 1, 2, 3\}$ is the product of above three t-distributions. Here $P \sim T(\alpha,\mu,\sigma^2)$ denote that the variable P has a t-distribution with degrees of freedom α, location parameter μ and scale parameter σ; i.e., $(P-\mu)/\sigma$ has the student t-distribution with degrees of freedom α. Furthermore, the generalized bayes estimator of $\psi = \theta_1\theta_2\theta_3$ is $\hat{\psi}_1 = \bar{x}_1\bar{x}_2\bar{x}_3$.

For one population, the fact that the marginal posterior distribution of θ_i given $(\bar{x}_i,s_i)$ has a T distribution is a known result (see Ref. 4). Rest of the proof follows immediately due to the fact that x_i values are independent and priors on different populations are independent.

In order to obtain a $100(1-\alpha)\%$ credible region, we simulate 20 000 values from the posterior density of $\psi(=\theta_1\theta_2\theta_3)$ given $\{(\bar{x}_i,s_i);\ i = 1, 2, 3\}$ and then take the $(\alpha/2)$th percentile of the simulated density as the lower limit of the credible region and $(1-\alpha/2)$th percentile as the upper limit of the credible region. This can be done quite easily, since the posterior distribution of θ_i given $(\bar{x}_i,s_i)$ is a t-distribution. Notice here, even though the posterior density of θ_i given $(\bar{x}_i,s_i)$ is a symmetric density, the posterior density of the product $\psi = \theta_1\theta_2\theta_3$ is no longer a symmetric density nor a t-distribution and therefore the credible region calculated this way may not produce the shortest possible credible region.

A computer program to calculate confidence interval is given in the appendix (program 1). This program is written in Fortran and uses IMSL libraries. We run the program for three different cases.

Example 1. The data for this example is from Yfantis and Flatman [3]. Here all three sample sizes are the same, which is a requirement for their method. Actual population values: $\theta_1 = -4.0$, $\theta_2 = 0.2$, $\theta_3 = 100.0$, $\sigma_1 = 2.0$, $\sigma_2 = 2.0$, $\sigma_3 = 10.0$. Sample sizes: $n_1 = 64$, $n_2 = 64$, $n_3 = 64$. Generated sample values: $\bar{x}_1 = -4.1467$, $\bar{x}_2 = 0.3389$, $\bar{x}_3 = 101.5108$, $s_1 = 1.9968$, $s_2 = 1.9537$, $s_3 = 10.6592$. Estimated value for the product $\theta_1\theta_2\theta_3$: $\hat{\psi}_1 = -142.68$. 95% Confidence interval for product: (−350.42, 61.90).

Notice here that $\hat{\psi}_0 = -164.26$ and 95% confidence interval proposed by Yfantis and Flatman [3] is (−433.74, 75.71). As expected, confidence interval based on $\hat{\psi}_1$ is shorter than the confidence interval based on $\hat{\psi}_0$.

For the other examples, we generate random samples from known populations and then using these random samples we estimate the product of three means. Results are given in the following three examples.

Example 2. Actual population values: $\theta_1 = 10.0$, $\theta_2 = 4.0$, $\theta_3 = 5.0$, $\sigma_1 = 3.0$, $\sigma_2 = 2.0$, $\sigma_3 = 2.0$. Sample sizes: $n_1 = 15$, $n_2 = 20$, $n_3 = 25$. Generated sample values: $\bar{x}_1 = 11.3825$, $\bar{x}_2 = 3.7484$, $\bar{x}_3 = 4.9894$, $s_1 = 2.3989$, $s_2 = 1.5776$, $s_3 = 2.0206$. Estimated value for the product $\theta_1\theta_2\theta_3$: $\hat{\psi}_1 = 212.88$. 95% Confidence interval for product: (156.01, 276.70).

Example 3. Actual population values: $\theta_1 = 10.0$, $\theta_2 = 4.0$, $\theta_3 = 5.0$, $\sigma_1 = 2.0$, $\sigma_2 = 1.0$, $\sigma_3 = 2.0$. Sample sizes: $n_1 = 40$, $n_2 = 20$, $n_3 = 30$. Generated sample values: $\bar{x}_1 = 9.9675$, $\bar{x}_2 = 3.9202$, $\bar{x}_3 = 5.0929$, $s_1 = 2.0474$, $s_2 = 1.3846$, $s_3 = 2.1317$. Estimated value for the product $\theta_1\theta_2\theta_3$: $\hat{\psi}_1 = 199.01$. 95% Confidence interval for product: (154.31, 247.87).

Method 2

Suppose the prior density of θ_i and σ_i^2 has the form

$$\pi_i(\theta_i,\sigma_i^2) = \pi_{1i}(\theta_i/\sigma_i^2)\pi_{2i}(\sigma_i^2) \tag{3}$$

where $\pi_{1i}(\theta_i/\sigma_i^2)$ is a normal $N(\mu_i,\tau_i\sigma_i^2)$ density and $\pi_{2i}(\sigma_i^2)$ is an inverted gamma $IG(\alpha_i,\beta_i)$ density. The values μ_i, τ_i, α_i and β_i are assumed to be known. Here the prior knowledge of θ_i is depend on σ_i^2. This joint prior density is from a conjugate family for the distribution of X_i.

Theorem 2

The marginal posterior distribution of θ_i given $(\bar{x}_i,s_i)$ is

$$\theta_i/(\bar{x}_i,s_i) \sim T\Big\{2\alpha_i + n_i - 1,\ (\mu_i + n_i\tau_i\bar{x}_i)/(n_i\tau_i + 1),\ \big[(\tau_i^{-1} + n_i)\times(\alpha_i + (n_i - 1)/2)\beta_i'\big]^{-1}\Big\} \tag{4}$$

where

$$\beta_i' = \Big[\beta_i^{-1} + (n_i - 1)s_i^2/2 + n_i(\bar{x}_i - \mu_i)^2/(2 + 2n_i\tau_i)\Big]^{-1}$$

The joint posterior distribution of $(\theta_1,\theta_2,\theta_3)$ given $\{(\bar{x}_i,s_i);\ i = 1, 2, 3\}$ is the product of the above three t-distributions and the Bayes estimate of the product $\psi = \theta_1\theta_2\theta_3$ is

$$\hat{\psi}_2 = \prod_{i=1}^{i=3}(\mu_i + n_i\tau_i\bar{x}_i)/(n_i\tau_i + 1)$$

The result in Eqn. 4 is given by Berger [4]. The rest of the theorem follows similar to the Theorem 1.

Similar to the method 1, we simulate the posterior density of $\psi = \theta_1\theta_2\theta_3$ given $\{(\bar{x}_i,s_i);\ i = 1, 2, 3\}$ and use it to get $100(1-\alpha)\%$ credible region. This computer program is given in the Appendix (Program 2). Given the prior parameter values and sample values, this program calculate the Bayesian estimate and Bayesian credible region for ψ. Again, this credible region may not be the shortest possible answer since the posterior density is not a symmetric distribution. We use this method on three simulated data sets and their results are given in the following two examples.

Example 4. Prior parameter values: $\alpha_1 = 2.20$, $\alpha_2 = 2.20$, $\alpha_3 = 2.20$, $\beta_1 = 0.10$, $\beta_2 = 0.05$, $\beta_3 = 0.01$, $\mu_1 = 10.0$, $\mu_2 = 20.0$, $\mu_3 = 100.0$, $\tau_1 = 0.25$, $\tau_2 = 0.25$, $\tau_3 = 0.25$. Generated population val-

ues: $\theta_1 = 10.16$, $\theta_2 = 17.94$, $\theta_3 = 102.38$, $\sigma_1 = 3.47$, $\sigma_2 = 3.98$, $\sigma_3 = 7.58$. Sample sizes: $n_1 = 8$, $n_2 = 10$, $n_3 = 12$. Generated sample values: $\bar{x}_1 = 7.88$, $\bar{x}_2 = 20.42$, $\bar{x}_3 = 100.07$, $s_1 = 3.76$, $s_2 = 4.51$, $s_3 = 6.94$.

(a) Using Method 2 (using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_2 =$ 17435.9. 95% Confidence interval for product, (12824.3, 22438.3).

(b) Using Method 1 (without using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_1 = 16096.3$. 95% Confidence interval for product, (9475.95, 23237.00).

In this example sizes are very small and the confidence interval produced by Method 2 (Bayesian method) is much shorter than the confidence interval produced by Method 1 (the method without prior knowledge). If the accurate prior knowledge is available, this is true most of the time, in particular when the sample sizes are very small and the population variances are large.

Example 5. Prior parameter values: $\alpha_1 = 2.20$, $\alpha_2 = 2.20$, $\alpha_3 = 2.20$, $\beta_1 = 0.10$, $\beta_2 = 0.05$, $\beta_3 = 0.01$, $\mu_1 = 10.0$, $\mu_2 = 20.0$, $\mu_3 = 100.0$, $\tau_1 = 0.25$, $\tau_2 = 0.25$, $\tau_3 = 0.25$. Generated population values: $\theta_1 = 10.81$, $\theta_2 = 22.75$, $\theta_3 = 100.89$, $\sigma_1 = 2.86$, $\sigma_2 = 3.08$, $\sigma_3 = 6.16$. Sample sizes: $n_1 = 50$, $n_2 = 55$, $n_3 = 60$. Generated sample values: $\bar{x}_1 = 10.82$, $\bar{x}_2 = 23.19$, $\bar{x}_3 = 100.54$, $s_1 = 2.74$, $s_2 = 3.07$, $s_3 = 7.04$.

(a) Using Method 2 (using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_2 =$ 24833.2. 95% Confidence interval for product, (22863.9, 26773.6).

(b) Using Method 1 (without using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_1 = 25216.2$. 95% Confidence interval for product, (23141.8, 27322.7).

Method 3

In this method we assume that the population variances σ_1^2, σ_2^2 and σ_3^2 are known or the sample sizes are large enough so that unknown σ_i could be approximately estimated by the sample standard deviation s_i. Consider a normal prior $N(\mu_i,\tau_i^2)$ on each θ_i; $i = 1, 2, 3$. The values μ_i, τ_i are assumed to be known.

Theorem 3

The marginal posterior distribution of θ_i given $(\bar{x}_i,s_i)$ is

$$\theta_i/(\bar{x}_i,s_i) \sim N\left(\frac{n_i\bar{x}_i/\sigma_i^2 + \mu_i/\tau_i^2}{n_i/\sigma_i^2 + 1/\tau_i^2}, \frac{1}{n_i/\sigma_i^2 + 1/\tau_i^2}\right) \quad (5)$$

The joint posterior distribution of $(\theta_1,\theta_2,\theta_3)$ given $\{(\bar{x}_i,s_i); i = 1, 2, 3\}$ is the product of the above three normal distributions and the bayes estimate of the product $\psi = \theta_1\theta_2\theta_3$ is

$$\hat{\psi} = \prod_{i=1}^{i=3} \frac{n_i\bar{x}_i/\sigma_i^2 + \mu_i/\tau_i^2}{n_i/\sigma_i^2 + 1/\tau_i^2} \quad (6)$$

Credible regions for ψ can be found by simulating the posterior distribution of ψ.

The posterior distribution in the Eqn. 5, is a well known result [4]. The rest of the theorem follows since x_i values are independent and priors on different population are independent.

A simulation program to calculate $100(1-\alpha)\%$ credible region is given in the Appendix (Program 3). Given the prior parameter values and sample values, this program calculate the Bayesian estimate and Bayesian credible region for ψ. Again, this credible region may not be the shortest possible region since the posterior density is not a symmetric distribution. We run this program for three different cases. Their results are given in the following two examples.

Example 6. Prior parameter values: $\sigma_1 = 3.0$, $\sigma_2 = 5.0$, $\sigma_3 = 4.0$, $\mu_1 = 10.0$, $\mu_2 = 20.0$, $\mu_3 = 15.0$, $\tau_1 = 2.0$, $\tau_2 = 1.0$, $\tau_3 = 2.0$. Generated population values: $\theta_1 = 10.41$, $\theta_2 = 18.03$, $\theta_3 = 16.88$. Sample sizes: $n_1 = 12$, $n_2 = 15$, $n_3 = 16$. Generated sample values: $\bar{x}_1 = 9.75$, $\bar{x}_2 = 15.46$, $\bar{x}_3 = 16.21$, $s_1 = 3.08$, $s_2 = 4.70$, $s_3 = 3.72$.

(a) Method 3 (using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_3 = 2860.42$. 95% Confidence interval for product, (2291.99, 3504.70).

(b) Method 1 (without using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_1 =$ 2443.56. 95% Confidence interval for product, (1780.18, 3192.22).

Example 7. This is an example with large samples. In order to use Method 3, if the prior values for σ_1, σ_2 and σ_3 are not available, one can use sample standard deviations to estimate the population standard deviations (σ). Estimators and confidence intervals based on both cases (σ known, and σ unknown) are given in this example. Prior parameter values: $\sigma_1 = 4.0$, $\sigma_2 = 10.0$, $\sigma_3 = 10.0$, $\mu_1 = 10.0$, $\mu_2 = 50.0$, $\mu_3 = 40.0$, $\tau_1 = 3.0$, $\tau_2 = 2.0$, $\tau_3 = 2.0$. Generated population values: $\theta_1 = 6.23$, $\theta_2 = 48.55$, $\theta_3 = 44.37$. Sample sizes: $n_1 = 50$, $n_2 = 55$, $n_3 = 60$. Generated sample values: $\bar{x}_1 = 6.70$, $\bar{x}_2 = 49.56$, $\bar{x}_3 = 45.90$, $s_1 = 3.80$, $s_2 = 10.95$, $s_3 = 10.71$.

(a) Method 3 (using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_3 = 14962.2$. 95% Confidence interval for product, (12420.4, 17527.7).

(b) Method 3 (using prior knowledge, except for σ): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_3 = 13808.0$. 95% Confidence interval for product, (11274.9, 16595.8).

(c) Method 1 (without using prior knowledge): estimated value for the product $\theta_1\theta_2\theta_3$, $\hat{\psi}_1 = 15249.6$. 95% Confidence interval for product, (12555.9, 18117.1).

DISCUSSION

In statistical inference, Bayesian methods are known to produce better results whenever the correct prior knowledge is available. In particular, if you are using small size samples or moderate size samples with large population variances, it is advantageous to use prior information if available. If no prior information is available, one need to use the estimator $\hat{\psi}_1$ and confidence intervals based on $\hat{\psi}_1$ instead of the estimator $\hat{\psi}_0$. Not only the estimator $\hat{\psi}_1$ has a smaller variance than $\hat{\psi}_0$, but also it has certain optimal properties since it is the generalized Bayes estimator with respect to the noninformative prior. Furthermore equal size samples are not a requirement for $\hat{\psi}_1$ which enhances the use of $\hat{\psi}_1$ over $\hat{\psi}_0$.

APPENDIX

```
ccccccccccccccccccc
c     PROGRAM 1      c
ccccccccccccccccccc

c Given sample means and sample standard deviations, this program calculate
c the generalized Bayes estimator and a 100(1-a)% credible region
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      real   r1,r2,r3,x1,x2,x3,s1,s2,s3,alpha
      print*, 'Enter sample sizes respectively '
      read*, r1,r2,r3
      print*, 'Enter sample means respectively'
      read*, x1,x2,x3
      print*, 'Enter sample standard deviations respectively'
      read*, s1,s2,s3
      print*, 'Enter confidence coefficient, i.e. if you need 95% C.I.
     &   enter .05 etc. '
      read*, alpha
      print*, 'Program is running, please wait..'
      niter=20000
      call cutoff(alpha,niter,clim1,clim2)
      print*, '    Estimator for the product is  ',x1*x2*x3
      print*, 100*(1-alpha),'%', '  confidence interval is      ('
     &    , clim1, ',' , clim2, ')'
      stop
      end
c
      subroutine cutoff(alpha,niter, clim1,clim2)
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      real estdis(20000)
      call generate(niter,estdis)
```

```
      call svrgn(niter,estdis,estdis)
      l1= int( (alpha/2.)* real(niter) )
      clim1= ( estdis(l1)+estdis(l1+1) )/2.
      l2= int( (1.-alpha/2.)*real(niter) )
      clim2= ( estdis(l2)+ estdis(l2+1) )/2.
      return
      end
c
      subroutine generate(niter,estdis)
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      real rsamp1(20000),rsamp2(20000),rsamp3(20000),estdis(20000)
      call rnun(niter,rsamp1)
      call rnun(niter,rsamp2)
      call rnun(niter,rsamp3)
      do 1212 ite=1,niter
      rsamp1(ite)= x1+ tin( rsamp1(ite), r1-1.)*s1/sqrt(r1)
      rsamp2(ite)= x2+ tin( rsamp2(ite), r2-1.)*s2/sqrt(r2)
      rsamp3(ite)= x3+ tin( rsamp3(ite), r3-1.)*s3/sqrt(r3)
      estdis(ite)=rsamp1(ite)*rsamp2(ite)*rsamp3(ite)
1212  continue
      return
      end
c End of the program.
ccccccccccccccccccccc
c     PROGRAM 2       c
ccccccccccccccccccccc
c Given sample means and sample standard deviationsi and prior information,
c this program  calculate bayesian estimator and a 100(1-a)% credible region
c for the product.
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      common rmu1,rmu2,rmu3,tao1,tao2,tao3,alp1,alp2,alp3,
     &    beta1,beta2,beta3
      real r1,r2,r3,x1,x2,x3,s1,s2,s3,rmu1,rmu2,rmu3,tao1,tao2,
     &     tao3,alp1,alp2,alp3,beta1,beta2,beta3
      print*, 'Enter sample sizes respectively'
      read*, r1,r2,r3
      print*, 'Enter sample means respectively'
      read*, x1,x2,x3
      print*, 'Enter sample standard deviations respectively'
      read*, s1,s2,s3
      print*, 'Enter confidence coefficient, i.e. if you need 95% C.I.
     &  enter .05  etc.'
      read*, alpha
      print*, 'Now enter prior parameters'
      print*, 'Enter prior means respectively, i.e. mu1,mu2,mu3'
      read*, rmu1,rmu2,rmu3
      print*, 'Enter tao values respectively'
      read*, tao1,tao2,tao3
      print*, 'Enter alpha values respectively'
      read*, alp1,alp2,alp3
      print*, 'Enter  beta values respectively'
      read*, beta1,beta2,beta3
      print*, 'Program is running, please wait...'
      niter=20000
      call cutoff(alpha,niter,clim1,clim2)
      print*,'    Estimator for the product is      ',stdest(r1,x1
     & ,rmu1,tao1)*stdest(r2,x2,rmu2,tao2)*stdest(r3,x3,rmu3,tao3)
      print*,  100*(1-alpha), '%   confidence interval is     (',
     &       clim1, ',', clim2, '  )'
      stop
      end
c
      subroutine cutoff(alpha,niter, clim1,clim2)
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      common rmu1,rmu2,rmu3,tao1,tao2,tao3,alp1,alp2,alp3,
     &    beta1,beta2,beta3
```

```
      real estdis(20000)
      call generate(niter,estdis)
      call svrgn(niter,estdis,estdis)
      l1= int( (alpha/2.)* real(niter) )
      clim1= ( estdis(l1)+estdis(l1+1) )/2.
      l2= int( (1.-alpha/2.)*real(niter) )
      clim2= ( estdis(l2)+ estdis(l2+1) )/2.
      return
      end
c
      subroutine generate(niter,estdis)
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      common rmu1,rmu2,rmu3,tao1,tao2,tao3,alp1,alp2,alp3,
     &    beta1,beta2,beta3
      real rsamp1(20000),rsamp2(20000),rsamp3(20000),estdis(20000)
      betapr1=1./ (  1./beta1  + (r1-1.)*s1**2/2. +
     &  r1*(x1-rmu1)**2/( 2.*(1.+r1*tao1))   )
      betapr2=1./ (  1./beta2  + (r2-1.)*s2**2/2. +
     &  r2*(x2-rmu2)**2/( 2.*(1.+r2*tao2))   )
      betapr3=1./ (  1./beta3  + (r3-1.)*s3**2/2. +
     &  r3*(x3-rmu3)**2/( 2.*(1.+r3*tao3))   )
      call rnun(niter,rsamp1)
      call rnun(niter,rsamp2)
      call rnun(niter,rsamp3)
      do 1212 ite=1,niter
      rsamp1(ite)=stdest(r1,x1,rmu1,tao1)
     &  + tin(rsamp1(ite), 2.*alp1+r1-1.)
     &  * ( (1./tao1+r1)*(alp1+(r1-1.)/2.)*betapr1  )**(-.5)
      rsamp2(ite)=stdest(r2,x2,rmu2,tao2)
     &  + tin(rsamp2(ite), 2.*alp2+r2-1.)
     &  * ( (1./tao2+r2)*(alp2+(r2-1.)/2.)*betapr2  )**(-.5)
      rsamp3(ite)=stdest(r3,x3,rmu3,tao3)
     &  + tin(rsamp3(ite), 2.*alp3+r3-1.)
     &  * ( (1./tao3+r3)*(alp3+(r3-1.)/2.)*betapr3  )**(-.5)
      estdis(ite)=rsamp1(ite)*rsamp2(ite)*rsamp3(ite)
1212  continue
      return
      end
c%%%%%
      real function stdest(ri,xx1,rmu,tao)
      stdest= (rmu+ri*tao*xx1)/(ri*tao+1.)
      return
      end
c End of the program.



ccccccccccccccccccccccccc
c      PROGRAM 3           c
ccccccccccccccccccccccccc

c Given sample means, sample standard deviations and prior infomation,
c this program calculate bayesian estimators and 100(1-a)% credible
c region using a normal prior.
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      common rmu1,rmu2,rmu3,tao1,tao2,tao3,sigma1,sigma2,sigma3
      print*, 'Enter sample sizes respectively'
      read*, r1,r2,r3
      print*, 'Enter sample means respectively'
      read*, x1,x2,x3
      print*, 'Enter sample standard deviations respectively'
      read*, s1,s2,s3
      print*, 'Enter confidence coefficient, i.e. if you need 95% C.I.
     &  enter .05  etc.'
      read*, alpha
      print*, 'Now enter prior parameters'
```

```
      print*, 'Enter prior means respectively, i.e. mu1,mu2,mu3'
      read*, rmu1,rmu2,rmu3
      print*, 'Enter tao values respectively'
      read*, tao1,tao2,tao3
      print*, 'Enter sigma values respectively'
      read*, sigma1,sigma2,sigma3
      print*, 'Program is running, please wait...'
      niter=20000
      call cutoff(alpha,niter,clim1,clim2)
      print*,'    Estimator for the product is      ',
     & stdest(r1,x1,rmu1,sigma1,tao1)*stdest(r2,x2,rmu2,sigma2,tao2)
     &    *stdest(r3,x3,rmu3,sigma3,tao3)
      print*,  100*(1-alpha), '%   confidence interval is     (',
     &       clim1, ',', clim2, '  )'
      stop
      end
c
      subroutine cutoff(alpha,niter, clim1,clim2)
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      common rmu1,rmu2,rmu3,tao1,tao2,tao3,sigma1,sigma2,sigma3
      real estdis(20000)
      call generate(niter,estdis)
      call svrgn(niter,estdis,estdis)
      l1= int( (alpha/2.)* real(niter) )
      clim1= ( estdis(l1)+estdis(l1+1) )/2.
      l2= int( (1.-alpha/2.)*real(niter) )
      clim2= ( estdis(l2)+ estdis(l2+1) )/2.
      return
      end
c
      subroutine generate(niter,estdis)
      common r1,r2,r3,x1,x2,x3,s1,s2,s3
      common rmu1,rmu2,rmu3,tao1,tao2,tao3,sigma1,sigma2,sigma3
      real rsamp1(20000),rsamp2(20000),rsamp3(20000),estdis(20000)
      call rnnor(niter,rsamp1)
      call rnnor(niter,rsamp2)
      call rnnor(niter,rsamp3)
      do 1212 ite=1,niter
      rsamp1(ite)= stdest(r1,x1,rmu1,sigma1,tao1)+ rsamp1(ite)
     &  / sqrt( r1/sigma1**2 + 1./tao1**2 )
      rsamp2(ite)= stdest(r2,x2,rmu2,sigma2,tao2)+ rsamp2(ite)
     &  / sqrt( r2/sigma2**2 + 1./tao2**2 )
      rsamp3(ite)= stdest(r3,x3,rmu3,sigma3,tao3)+ rsamp3(ite)
     &  / sqrt( r3/sigma3**2 + 1./tao3**2 )
      estdis(ite)=rsamp1(ite)*rsamp2(ite)*rsamp3(ite)
1212  continue
      return
      end
c
      real function stdest(ri,xx1,rmu,sig,tao)
      stdest= ( ri*xx1/sig**2 + rmu/tao**2 )/
     &    ( ri/sig**2 + 1./tao**2 )
      return
      end
c End of the program.
```


REFERENCES

1 S. Hogg and A. Craig, Introduction to Mathematical Statistics, Macmillian, New York, 1978.

2 United States Environmental Protection Agency, EPA 540189-002, Vol. 1. 1989, Chap. 6.

3 E. Yfantis and G.T. Flatman, J. Chemom., 5 (1991) 309.

4 J. Berger, Statistical Decision Theory, Springer-Verlag, New York, 1985.

ANALYTICA CHIMICA ACTA, VOL. 277 (1993)

AUTHOR INDEX

Ananda, M.M.A.
—, Singh, A.K. and Flatman, G.T.
Bayesian confidence intervals for the product of three normal means 503
Ananda, M.M.A., see Singh, A.K. 255
Audunsson, G., see Jönsson, J.Å. 9

Barreto, W.J., see Zaia, D.A.M. 89
Baum, E.J.
Receptor models to study groundwater contamination 421
Beasy, M.A., see White, R.L. 333
Blank, T.B.
— and Brown, S.D.
Data processing using neural networks 273
Bond, A.M.
—, Pfund, B.V. and Newman, O.M.G.
Polarographic determination of total iron, iron(II) and iron(III) in zinc plant electrolyte 145
Borgen, O.S., see Ukkelberg, Å. 489
Bos, A.
—, Bos, M. and Van der Linden, W.E.
Artificial neural networks as a multivariate calibration tool: modeling the Fe–Cr–Ni system in x-ray fluorescence spectroscopy 289
Bos, M., see Bos, A. 289
Bramanti, E., see Scarano, G. 137
Brodmeier, T.
—, Gloor, A., Cadisch, M., Bürgin, R. and Pretsch, E.
Hypermedia tools for the interpretation of mass spectra 297
Brown, S.D., see Blank, T.B. 273
Bürgin, R., see Brodmeier, T. 297
Buydens, L., see Wehrens, R. 313

Cadisch, M., see Brodmeier, T. 297
Carr, P.W., see Poe, R.B. 223
Chen, Z., see Wang, J. 153
Cheng, M.-D., see Gao, N. 369
Chiba, Y., see Sato, K. 61
Ci, Y.-X.
—, Tie, J.-K., Yao, F.-J., Liu, Z.-L., Lin, S. and Zheng, W.-Q.
Catalytic behaviour of iron(II)–oxime complexes in the chemiluminescence reaction of luminol with hydrogen peroxide 67
Clark, D.S., see White, R.L. 333
Cullen, M.
— and Mearns, J.M.P.
Coulometric determination of arsenic in gallium arsenide crystal wafers 113

Danielson, N.D., see Holeman, J.A. 55
Dattner, S., see Spiegelman, C.H. 347
Deming, S.N., see Olivero, R.A. 441
Dombek, V., see Praus, P. 97
Doornbos, D.A., see Duineveld, C.A.A. 455
Duineveld, C.A.A.
—, Smilde, A.K. and Doornbos, D.A.
Designs for mixture and process variables applied in tablet formulations 455
Dunn, III, W.J., see Swain, D. 305

Eckert, C.A., see Poe, R.B. 223
Endo, A.S., see Zaia, D.A.M. 89

Flatman, G.T., see Ananda, M.M.A. 503
Frenzel, W.
—, Schepers, D. and Schulze, G.
Simultaneous ion chromatographic determination of anions and cations by series conductivity and flame photometric detection 103

Gao, N.
—, Cheng, M.-D. and Hopke, P.K.
Potential source contribution function analysis and source apportionment of sulfur species measured at Rubidoux, CA during the Southern California Air Quality Study, 1987 369
Garner, F.C., see Miah, M.J. 431
Gleser, L.J.
— and Yang, H.
Empirical Bayes estimation in factor analysis for aerosol mass apportionment 405
Gloor, A., see Brodmeier, T. 297
Goeringer, D.E., see Van Berkel, G.J. 41
Griffin, K.J., see Moen, M.O. 477
Grund, D.W., see White, R.L. 333
Guan, J.-S., see Yang, X.-J. 157

Hait, M.J., see Poe, R.B. 223
Harrington, P. de B.
— and Pack, B.W.
FLIN: Fuzzy Linear Interpolating Network 189

Hayashi, Y.
— and Matsuda, R.
Information theory of chromatography and titration 325
Hieftje, G.M., see Madrid, Y. 1
Hill, N.C., see Limbach, P.A. 31
Holeman, J.A.
— and Danielson, N.D.
Chemiluminescence reaction of thiazide compounds with tris(2,2′-bipyridine)ruthenium(III) 55
Hopke, P.K., see Gao, N. 369

Ishii, H., see Odashima, T. 79
Iwata, A.
—, Yamanaka, C., Nakashima, N. and Izawa, Y.
On-line real-time detection method for trace Cl^- in aqueous solution 25
Izawa, Y., see Iwata, A. 25

Jin, Q., see Madrid, Y. 1
Jonsson, J., see Wold, S. 239
Jönsson, J.Å.
—, Lövkvist, P., Audunsson, G. and Nilvé, G.
Mass transfer kinetics for analytical enrichment and sample preparation using supported liquid membranes in a flow system with stagnant acceptor liquid 9

Kalantar, A.H., see Moen, M.O. 477
Kateman, G.
— and Smits, J.R.M.
Colored information from a black box? Validation and evaluation of neural networks 179
Kateman, G., see Wehrens, R. 313
Kaufmann, P.
Prediction of mixture composition by chromatographic characterization, multivariate classification and partial least-squares regression, a comparison of methods 467
Kim, H.S., see Limbach, P.A. 31
Kokkinidis, G., see Papanastasiou, G. 119
Kowalski, B., see Seasholtz, M.B. 165
Krock, K.A.
— and Wilkins, C.L.
Environmental applications of combined multidimensional gas chromatography–infrared–mass spectrometry 381

Laing, G.A., see Miah, M.J. 431
Lavine, B.K.
—, Stine, A. and Mayfield, H.T.
Gas chromatography–pattern recognition techniques in pollution monitoring 357
Li, M., see Yu, B. 199
Li, Z., see Yu, B. 199
Limbach, P.A.
—, Kim, H.S., Hill, N.C. and Marshall, A.G.
Fast neutral beam Fourier transform ion cyclotron resonance mass spectrometry for analysis of insulating and conductive materials 31
Lin, S., see Ci, Y.-X. 67
Liu, A., see Yu, B. 199
Liu, Z.-L., see Ci, Y.-X. 67
Lövkvist, P., see Jönsson, J.Å. 9
Lu, Y., see Wang, J. 153
Lucasius, C., see Wehrens, R. 313
Luo, Y., see Wang, J. 153

Madrid, Y.
—, Wu, M., Jin, Q. and Hieftje, G.M.
Evaluation of flow-injection techniques for microwave plasma torch atomic emission spectrometry 1
Marshall, A.G., see Limbach, P.A. 31
Matsuda, R., see Hayashi, Y. 325
Mayfield, H.T., see Lavine, B.K. 357
Mearns, J.M.P., see Cullen, M. 113
Meloun, M., see Militký, J. 215, 267
Miah, M.J.
—, Garner, F.C., Stapanian, M.A. and Laing, G.A.
Optimal frequency of quality control analyses in environmental measurement 431
Militký, J.
— and Meloun, M.
Some graphical aids for univariate exploratory data analysis 215
— and Meloun, M.
Use of the mean quadratic error of prediction for the construction of biased linear models 267
Moen, M.O.
—, Griffin, K.J. and Kalantar, A.H.
Simple regression and outlier detection using the median method 477

Nakashima, N., see Iwata, A. 25
Newman, O.M.G., see Bond, A.M. 145
Nilvé, G., see Jönsson, J.Å. 9

Odashima, T.
— and Ishii, H.
Synthesis and properties of hydrazones from 3- and/or 5-nitro-2-pyridylhydrazine and heterocyclic aldehydes, characterization of their complexes and extraction–spectrophotometric determination of traces of nickel with 2-pyridinecarbaldehyde 3,5-dinitro-2-pyridylhydrazone 79
Ogawa, K., see Watarai, H. 73
Olivero, R.A.
—, Seshadri, S. and Deming, S.N.
Development of an expert system for selection of experimental designs 441

Pack, B.W., see Harrington, P. de B. 189
Pan, S., see Wang, J. 153
Pan, Z., see Yu, B. 199
Papanastasiou, G.
—, Ziogas, I. and Kokkinidis, G.
Simultaneous determination of equivalence volumes and thermodynamic acid dissociation constants from data for the acidic region of potentiometric titration curves 119
Penlidis, A., see Phatak, A. 495
Pfund, B.V., see Bond, A.M. 145

Phatak, A.
—, Reilly, P.M. and Penlidis, A.
An approach to interval estimation in partial least squares regression 495
Poe, R.B.
—, Rutan, S.C., Hait, M.J., Eckert, C.A. and Carr, P.W.
Developing models for infinite dilution activity coefficients using factor analysis methods 223
Praus, P.
— and Dombek, V.
Separation of chlorophenols by capillary isotachophoresis 97
Pretsch, E., see Brodmeier, T. 297

Rännar, S., see Wold, S. 239
Reilly, P.M., see Phatak, A. 495
Rutan, S.C., see Poe, R.B. 223

Sandberg, M., see Wold, S. 239
Santos, N.J., see Zaia, D.A.M. 89
Sato, K.
—, Chiba, Y. and Tanaka, S.
Selective determination of L-ascorbic acid by a chemiluminescence method with ascorbate oxidase 61
Scarano, G.
— and Bramanti, E.
Voltammetric behaviour of marine hydrophobic copper complexes: effect of adsorption processes at a mercury electrode 137
Schepers, D., see Frenzel, W. 103
Schulze, G., see Frenzel, W. 103
Seasholtz, M.B.
— and Kowalski, B.
The parsimony principle applied to multivariate calibration 165
Seshadri, S., see Olivero, R.A. 441
Shen, T.-J., see Yang, X.-J. 157
Shi, L., see Yu, B. 199
Singh, A.
Multivariate decision and detection limits 205
Singh, A., see Singh, A.K. 473
Singh, A.K.
— and Singh, A.
Bayesian slippage test for detection of outlying sub-samples 473
—, Ananda, M.M.A. and Sparks, A.R.
Superfund site characterization using non-parametric variogram modeling 255
Singh, A.K., see Ananda, M.M.A. 503
Sjöström, M., see Wold, S. 239
Smilde, A.K., see Duineveld, C.A.A. 455
Smits, J.R.M., see Kateman, G. 179
Sparks, A.R., see Singh, A.K. 255
Spiegelman, C.H.
— and Dattner, S.
Applying and developing receptor models to the 1990 El Paso air data: a look at receptor modeling with uncharacterized sources and graphical diagnostics 347
Stapanian, M.A., see Miah, M.J. 431
Stine, A., see Lavine, B.K. 357
Suzuki, N., see Watarai, H. 73
Swain, D.
—, Dunn, III, W.J. and Talaat, R.E.
Pattern recognition studies of tandem mass spectra 305

Talaat, R.E., see Swain, D. 305
Tanaka, S., see Sato, K. 61
Tie, J.-K., see Ci, Y.-X. 67

Ukkelberg, Å.
— and Borgen, O.S.
Outlier detection by robust alternating regression 489

Van Berkel, G.J.
— and Goeringer, D.E.
Gas chromatography–tandem mass spectrometry implemented on a bench-top quadrupole ion trap-based instrument using random noise to effect collision-induced dissociation 41
Van der Linden, W.E., see Bos, A. 289
Vong, R.J.
Atmospheric chemometrics for identification of trace element sources in precipitation 389

Wang, J.
—, Luo, Y., Chen, Z., Lu, Y., Pan, S. and Wang, X.
Remote analysis of motorboat exhausts using Fourier transform infrared spectrometry 153
Wang, X., see Wang, J. 153
Watarai, H.
—, Ogawa, K. and Suzuki, N.
Formation of fluorescent complexes of Eu(III) and Sm(III) with β-diketones and trioctylphosphine oxide in oil–water microemulsions 73
Wehrens, R.
—, Lucasius, C., Buydens, L. and Kateman, G.
HIPS, a hybrid self-adapting expert system for nuclear magnetic resonance spectrum interpretation using genetic algorithms 313
Wentzell, P.D., see White, R.L. 333
White, R.L.
—, Wentzell, P.D., Beasy, M.A., Clark, D.S. and Grund, D.W.
Taxonomy of *Amanita* mushrooms by pattern recognition of amino acid chromatographic data 333
Wilkins, C.L., see Krock, K.A. 381
Wold, S.
—, Jonsson, J., Sjöström, M., Sandberg, M. and Rännar, S.
DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures 239
Wu, M., see Madrid, Y. 1

Yamanaka, C., see Iwata, A. 25
Yang, H., see Gleser, L.J. 405

Yang, X.-J.
—, Guan, J.-S. and Shen, T.-J.
Determination of trace zirconium and hafnium in high-purity scandium oxide by inductively coupled plasma atomic emission spectrometry and extraction chromatography 157
Yao, F.-J., see Ci, Y.-X. 67
Yu, B.
—, Li, M., Liu, A., Li, Z., Shi, L. and Pan, Z.
Some novel methods based on recursive optimal estimation. Applications to analytical chemistry 199
Zaia, D.A.M.
—, Barreto, W.J., Santos, N.J. and Endo, A.S.
Spectrophotometric method for the simultaneous determination of proteins and amino acids with *p*-benzoquinone 89
Zheng, W.-Q., see Ci, Y.-X. 67
Ziogas, I., see Papanastasiou, G. 119

PUBLICATION SCHEDULE FOR 1993

	S'92	O'92	N'92	D'92	J	F	M	A	M	J	J	A
Analytica Chimica Acta	267/1 267/2	268/1 268/2	269/1 269/2	270/1 270/2	271/1 271/2	272/1 272/2 273/1–2	274/1 274/2	275/1–2 276/1 276/2	277/1 277/2	278/1 278/2	279/1 279/2	280/1 280/2
Vibrational Spectroscopy		4/1			4/2		4/3	5/1		5/2		5/3

INFORMATION FOR AUTHORS

Manuscripts. The language of the journal is English. English linguistic improvement is provided as part of the normal editorial processing. Authors should submit three copies of the manuscript in clear double-spaced typing on one side of the paper only. *Vibrational Spectroscopy* also accepts papers in English only.

Abstract. All papers and reviews begin with an Abstract (50–250 words) which should comprise a factual account of the contents of the paper, with emphasis on new information.

Figures. Figures should be prepared in black waterproof drawing ink on drawing or tracing paper of the same size as that on which the manuscript is typed. One original (or sharp glossy print) and two photostat (or other) copies are required. Attention should be given to line thickness, lettering (which should be kept to a minimum) and spacing on axes of graphs, to ensure suitability for reduction in size on printing. Axes of a graph should be clearly labelled, along the axes, outside the graph itself. All figures should be numbered with Arabic numerals, and require descriptive legends which should be typed on a separate sheet of paper. Simple straight-line graphs are not acceptable, because they can readily be described in the text by means of an equation or a sentence. Claims of linearity should be supported by regression data that include slope, intercept, standard deviations of the slope and intercept, standard error and the number of data points; correlation coefficients are optional.

Photographs should be glossy prints and be as rich in contrast as possible; colour photographs cannot be accepted. Line diagrams are generally preferred to photographs of equipment.

Computer outputs for reproduction as figures must be good quality on blank paper, and should preferably be submitted as glossy prints.

Nomenclature, abbreviations and symbols. In general, the recommendations of the International Union of Pure and Applied Chemistry (IUPAC) should be followed, and attention should be given to the recommendations of the Analytical Chemistry Division in the journal *Pure and Applied Chemistry* (see also *IUPAC Compendium of Analytical Nomenclature, Definitive Rules*, 1987).

References. The references should be collected at the end of the paper, numbered in the order of their appearance in the text (*not* alphabetically) and typed on a separate sheet.

Reprints. Fifty reprints will be supplied free of charge. Additional reprints (minimum 100) can be ordered. An order form containing price quotations will be sent to the authors together with the proofs of their article.

Papers dealing with vibrational spectroscopy should be sent to: Dr J.G. Grasselli, 150 Greentree Road, Chagrin Falls, OH 44022, U.S.A. Telefax: (+1-216) 2473360 (Americas, Canada, Australia and New Zealand) or Dr J.H. van der Maas, Department of Analytical Molecule Spectrometry, Faculty of Chemistry, University of Utrecht, P.O. Box 80083, 3508 TB Utrecht, The Netherlands. Telefax: (+31-30) 518219 (all other countries).

0003-2670/93/$06.00

This issue is printed on acid-free paper.

PRINTED IN THE NETHERLANDS

Experimental Design: A Chemometric Approach

Second, Revised and Expanded Edition

by **S.N. Deming** *and* **S.L. Morgan**

Data Handling in Science and Technology Volume 11

Now available is the second edition of a book which has been described as *"...an exceptionally lucid, easy-to-read presentation... would be an excellent addition to the collection of every analytical chemist. I recommend it with great enthusiasm."* (Analytical Chemistry).

N.R. Draper reviewed the first edition in Publication of the International Statistical Institute *"...discussion is careful, sensible, amicable, and modern and can be recommended for the intended readership."*

The scope of the first edition has been revised, enlarged and expanded. Approximately 30% of the text is new. The book first introduces the reader to the fundamentals of experimental design. Systems theory, response surface concepts, and basic statistics serve as a basis for the further development of matrix least squares and hypothesis testing. The effects of different experimental designs and different models on the variance-covariance matrix and on the analysis of variance (ANOVA) are extensively discussed. Applications and advanced topics (such as confidence bands, rotatability, and confounding) complete the text. Numerous worked examples are presented.

The clear and practical approach adopted by the authors makes the book applicable to a wide audience. It will appeal particularly to those who still need to know efficient ways of carrying out experiments. It will also be an ideal text for advanced undergraduate and graduate students following courses in chemometrics, data acquisition and treatment, and design of experiments.

Contents:

1. System Theory.
2. Response Surfaces.
3. Basic Statistics.
4. One Experiment.
5. Two Experiments.
6. Hypothesis Testing.
7. The Variance-Covariance Matrix.
8. Three Experiments.
9. Analysis of Variance (ANOVA) for Linear Models.
10. An Example of Regression Analysis on Existing Data.
11. A Ten-Experiment Example.
12. Approximating a Region of a Multifactor Response Surface.
13. Confidence Intervals for Full Second-Order Polynomial Models.
14. Factorial-Based Designs.
15. Additional Multifactor Concepts and Experimental Designs.

Appendix A. Matrix Algebra.
Appendix B. Critical Values of t.
Appendix C. Critical Values of F, α=0.05.
Subject Index.

1993 416 pages
Price: US $ 177.25 / Dfl. 310.00
ISBN 0-444-89111-0

ORDER INFORMATION

For USA and Canada
ELSEVIER SCIENCE PUBLISHERS
Judy Weislogel
P.O. Box 945
Madison Square Station,
New York, NY 10160-0757
Tel: (212) 989 5800
Fax: (212) 633 3880

In all other countries
ELSEVIER SCIENCE PUBLISHERS
P.O. Box 211
1000 AE Amsterdam
The Netherlands
Tel: (+31-20) 5803 753
Fax: (+31-20) 5803 705

US$ prices are valid only for the USA & Canada and are subject to exchange rate fluctuations; in all other countries the Dutch guilder price (Dfl.) is definitive. Customers in the European Community should add the appropriate VAT rate applicable in their country to the price(s). Books are sent post-free if prepaid.

ELSEVIER
SCIENCE PUBLISHERS

0003-2670(19930528)277:2;1-3