JOURNAL OF CHEMOMETRICS
J. Chemometrics **12,** 357–364 (1998)

POSITIVE MATRIX FACTORIZATION APPLIED TO A CURVE RESOLUTION PROBLEM

YU-LONG XIE,[1] PHILIP K. HOPKE[1]* AND PENTTI PAATERO[2]
[1]*Department of Chemistry, Clarkson University, Potsdam, NY 13699-5810, U.S.A.*
[2]*Department of Physics, University of Helsinki, PO Box 9, FIN-00014, Finland*

SUMMARY

Positive matrix factorization (PMF) is a least squares approach to factor analysis which was originally developed for environmental data analysis and has been applied to several problems in resolving sources of environmental pollutants. PMF has been used as both a two-way and three-way data analysis tool. In this investigation, three-way data arrays were used to explore the ability of PMF in curve resolution. Pulsed gradient spin echo (PGSE) nuclear magnetic resonance (NMR) data were measured for spectral mixtures where the concentrations of the compounds decay exponentially. Three-way data arrays were constructed by packing different parts of the data from single experiments and were analyzed with three-way PMF to obtain the NMR spectra, decay profiles and the self-diffusion coefficients of constituents. © 1998 John Wiley & Sons, Ltd.

KEY WORDS: positive matrix factorization (PMF); curve resolution; PGSE spectra; self-diffusion coefficients

INTRODUCTION

Instruments that generate two-way data arrays are now common in the analytical laboratory. Depending on whether the measurements are made on single or multiple samples, two- or three-way data can be subsequently formed. Calibration and spectral resolution are major foci of data analysis methods in analytical chemistry.

In two-way cases, research has focused on the resolution of complex mixtures and the assessment of peak purity based on the interpretation of hyphenated chromatographic data.[1] A variety of algorithms, mostly based on principal component analysis (PCA), have been proposed and the evolutionary character of hyphenated chromatographic data has been utilized to alleviate or eliminate ambiguity.

For three-way data, parallel factor analysis (PARAFAC) is the most commonly used method.[2] The trilinear structure of a data array can provide an identifiable solution that does not contain the rotational freedom that exists in all two-way analyses.[3] For a particular case where there are only two 'slices' in one of the orders of the three-way data array, the factorization can be done by solving a rectangular generalized eigenvalue–eigenvector problem using the generalized rank annihilation method (GRAM).[4]

Similar mathematical treatments with different physical models appear in data analysis used in

* Correspondence to: P. K. Hopke, Department of Chemistry, Clarkson University, Potsdam, NY 13699-5810, U.S.A.
Contract/grant sponsor: National Science Foundation; Contract/grant number: ATM 9523731;
Contract/grant sponsor: Vieho, Yrjö and Kalle Väisälä Foundation.

CCC 0886–9383/98/060357–08 $17.50

Received 4 November 1997
Accepted 23 March 1998

other scientific fields. For example, PCA, PARAFAC and other factorization techniques are also used for receptor modeling of environmental data in order to identify and apportion pollutants into various sources.[5]

Recently, a new mathematical technique called positive matrix factorization (PMF) was developed.[6] This technique differs from PCA in that individual error estimates of each data point are utilized and non-negativity and other constraints on the factors are integrated into the analysis. PMF can be applied to both two-way and three-way data arrays. PMF has been used in the analysis of environmental data, and specifically for receptor modeling of pollutants, and there have been a number of successful applications.[7–10]

As a general mathematical tool, PMF should also be useful for mixture resolution problems as are other chemometrics techniques. A simulated spectroscopic-like example[11] and a data set of fluorescence spectrographic measurements[12] have been previously used to demonstrate the potential of PMF for curve resolution. It is the aim of this study to examine the use of three-way PMF for curve resolution of experimental data. The three-way data previously analyzed by Windig and Antalek with GRAM[13] have been employed to test PMF.

POSITIVE MATRIX FACTORIZATION

The PMF algorithm is described in detail elsewhere,[6] so only a brief outline will be given here. Supposing $\underline{\mathbf{X}}$ is an $m \times n \times q$ three-way data array which results from the linear combination of p intrinsic factors in the measurement system, the trilinear decomposition of $\underline{\mathbf{X}}$ can be modeled as

$$\underline{\mathbf{X}} = \mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C} + \underline{\mathbf{E}} \tag{1}$$

or

$$\underline{X}_{ijk} = \sum_{h=1}^{p} A_{ih} B_{jh} C_{kh} + E_{ijk} \quad (i = 1, \ldots, m;\ j = 1, \ldots, n;\ k = 1, \ldots, q;\ h = 1, \ldots, p) \tag{2}$$

where **A, B** and **C** are the resolved two-way conforming factors for each of the three modes, the symbol $\otimes$ represents the tensor product and $\underline{\mathbf{E}}$ contains the residuals.

PMF solves an optimal weighted least squares task by assigning more realistic weights for each data point X_{ijk}, $w_{ijk} = 1/\sigma_{ijk}^2$ where the values of σ_{ijk} are the uncertainties associated with the measurements. Typically, σ_{ijk} is the uncertainty in the measured value X_{ijk}.

The trilinear model can be solved by minimizing the objective function

$$Q(E) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{q} w_{ijk} E_{ijk}^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{q} \left(\frac{E_{ijk}}{\sigma_{ijk}} \right)^2 \tag{3}$$

As a default option, PMF uses a penalty function to constrain the factors to be non-negative, as most physical conforming factors are never negative. It also includes regularization of the factors.[6] The factors are solved by an iterative optimization of both factors in each step,[6] which makes the algorithm efficient. A comparison of PMF[12] has been made with direct trilinear decomposition (DTD)[14] and several other trilinear model programs that are based on alternating least squares (ALS) methods. PMF was found to be much faster than the other ALS methods tested. Although it was not as fast as DTD, DTD produced poor results for some ill-conditioned problems.

PGSE NMR DATA

The PGSE NMR data were described by Windig and Antalek.[13] PGSE NMR uses a pulsed magnetic field gradient which influences the signal intensity of the resonance from the components in solution.

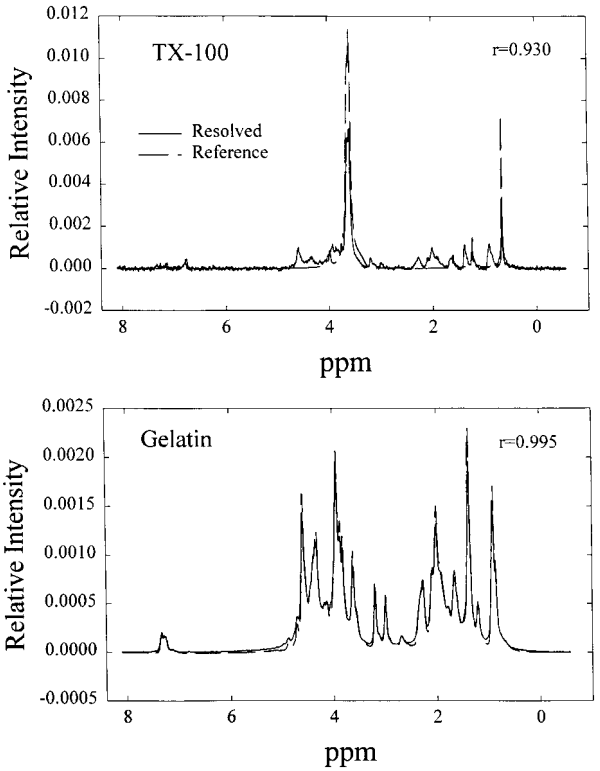

Figure 1. Two-factor PMF results for mixture 1: top, resolved and reference spectra of TX-100; bottom, resolved and reference spectra of gelatin.

The contribution of each of the components to the signal can be described by

$$\frac{A(i)}{A_0(i)} = \mathrm{e}^{-D(i)(\gamma g \delta)^2(\Delta - \delta/3)} \tag{4}$$

where $A(i)$ is the amplitude of component i, $D(i)$ is the self-diffusion coefficient of component i (m^2 s^{-1}), $A_0(i)$ is the amplitude with no magnetic gradients, γ is the gyromagnetic ratio of the ^1H nucleus (rad S^{-1} T^{-1}), g is the magnetic gradient strength (T), which is varied during the experiment so that the increments for subsequent g^2 values are the same, Δ is the diffusion time (s) and δ is the gradient width (s). The value of the term δ^2 $(\Delta - \delta/3)$ is given for a certain experiment.

The natural logarithm of the values of $A(i)$ is a linear function of g^2 where the slope is

$$s = -D(i)\gamma^2\delta^2(\Delta - \delta/3) \tag{5}$$

The slope can be calculated by simple regression. Since the experimental signal decays exponentially for every component and the decay is a function of the self-diffusion coefficient, PGSE NMR is used to determine these coefficients. The diffusion coefficient can then be calculated as

$$D(i) = -\frac{s}{7 \cdot 157 \times 10^{16} \times c} \tag{6}$$

In this equation the value of $7{\cdot}157 \times 10^{16}$ was substituted for γ^2, and c is the value of δ^2 $(\Delta - \delta/3)$,

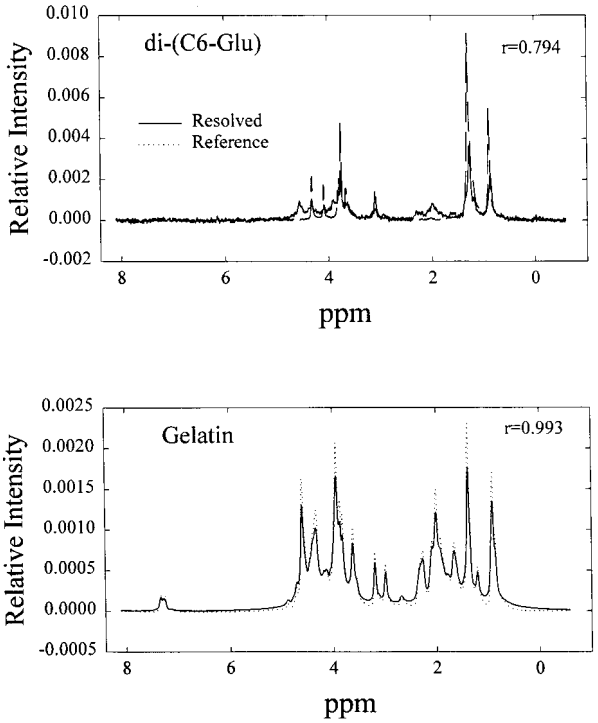

Figure 2. Two-factor PMF results for mixture 2: top, resolved and reference spectra of di-(C6-Glu); bottom, resolved and reference spectra of gelatin.

which is different for each experiment.

Three samples were analyzed in the PSGE experiments. The first sample is a mixture of 0·1% w/w TX-100, a non-ionic surfactant, and 5% w/w gelatin (mixture 1). The experiment resulted in 20 spectra with 4095 data points each (−0·58 to 8·1 ppm). The second sample is a mixture of 0·15% w/w di-(C6-Glu), a non-ionic surfactant, and 5% w/w gelatin (mixture 2). The experiment resulted in 20 spectra with 4095 data points each (−0·58 to 8·1 ppm). The third sample is a mixture of two components, 2-chloropropionic acid and 2-aminobenzothiozole, 0·94% w/w and 1·2% w/w respectively, in dimethyl sulfoxide-d_6 (DMSO) (mixture 3). Fifteen spectra were acquired with 6218 points each.

DATA ANALYSIS

According to Windig and Antalek,[13] the spectra at the beginning or/and the end of each data set are excluded in the data analysis because they are either caused by the self-diffusion of water or they deviate substantially from the other spectra. Thus only spectra 2–13, spectra 2–11 and spectra 2–15 were used for mixtures 1–3 respectively. Since the signal decays exponentially, two different parts of the data set can be used to create two data matrices, where the corresponding columns of these two matrices differ by only a scaling factor and the pure spectra and concentration profiles will have unit correlation coefficients. A three-way data array constructed by packing such two-way matrices together will therefore fit a trilinear model. Analogous to Windig and Antalek,[13] two data matrices for each data set were generated by taking the first to the $(n-1)$th and the second to the nth spectra of each

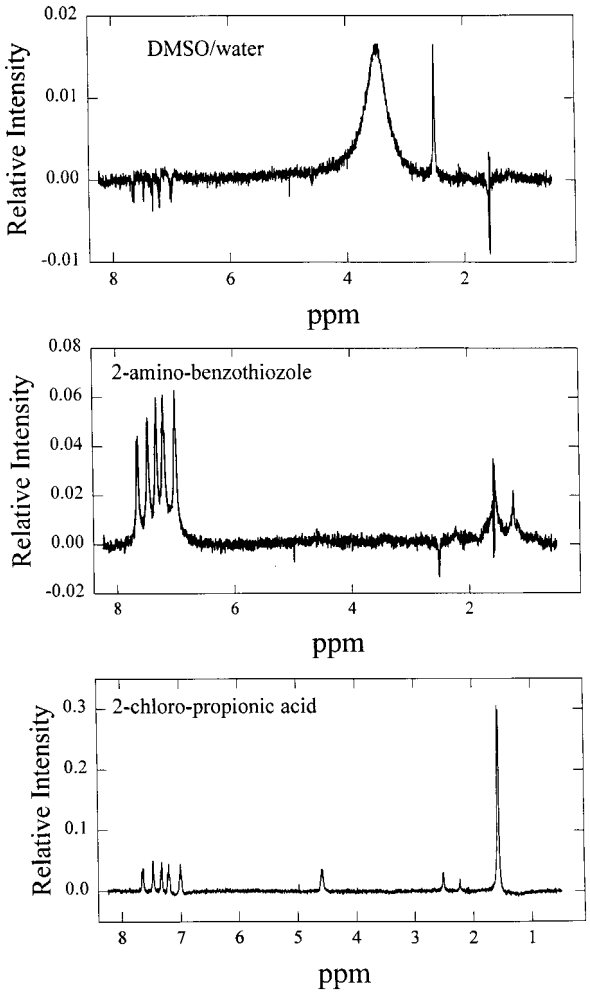

Figure 3. Three-factor PMF results for mixture 3: top, resolved spectrum of DMSO/water; middle, resolved spectrum of 2-aminobenzothiozole; bottom, resolved spectrum of 2-chloropropionic acid.

experimental data set respectively, and thus the three-way data arrays of these three mixtures have dimensions $4095 \times 11 \times 2$, $4095 \times 9 \times 2$ and $6218 \times 13 \times 2$ respectively, and PMF analyses were made.

RESULTS AND DISCUSSION

Error estimates for these measurements were not available with the data. To make full use of PMF, error estimates are needed for each data point, and thus they had to be estimated. PMF provides various models for error estimation. An absolute error plus a relative error proportional to the signal amplitude was assumed and the constants were chosen by trial and error. Examination of equation (3) suggests that the theoretical Q value should approach the number of data points in the data array if the error estimates are good approximations to the actual errors. Therefore PMF was run with different

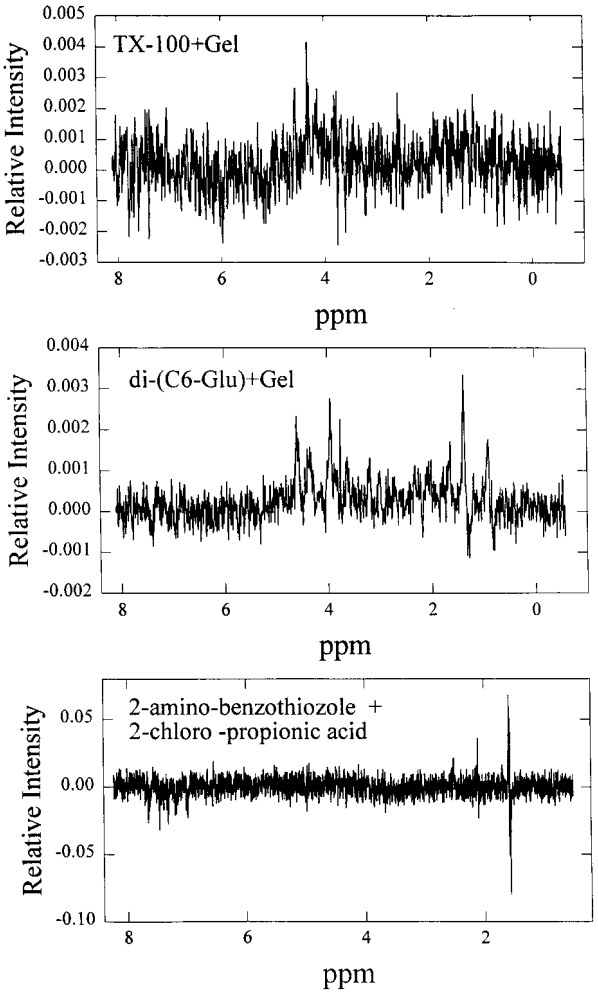

Figure 4. Extra factor when one factor more was used for mixtures: top, noise factor when three factors were used for mixture 1; middle, noise factor when three factors were used for mixture 2; bottom, noise factor when four factors were used for mixture 3.

constants representing absolute and relative errors and the PMF result with a Q value closest to the theoretical one was retained.

The original experimental data contain negative values that may be the result of instrumental noise. The default non-negativity option was deactivated in order not to introduce bias. The correct number of components of the data set was determined by observation of the resolved factors. It was found that when one factor more than necessary was extracted, the resulting spectrum clearly shows only noise.

It was found that two-factor models fit mixtures 1 and 2, but a three-factor model was needed for mixture 3. The resolved spectra of the components in the mixtures are shown in Figures 1–3 for mixtures 1–3 respectively. Models with one additional factor were calculated for each mixture. These extra factors are shown in Figure 4 and appear to consist only of noise, indicating that the choices for number of factors were correct.

Figure 1 shows the excellent agreement between the resolved and reference spectrum of gelatin.

Table 1. Self-diffusion coefficients and linearity of gelatin/TX-100 mixture

	Self-diffusion coefficient ($m^2\ s^{-1}$)				
	Slope[a]	Eigenvalue[a]	Direct[a]	PMF[b]	r[c]
Gelatin	$1{\cdot}32 \times 10^{-11}$	$1{\cdot}32 \times 10^{-11}$	$1{\cdot}32 \times 10^{-11}$	$1{\cdot}31 \times 10^{-11}/1{\cdot}32 \times 10^{-11}$	1·0000
TX-100	$9{\cdot}00 \times 10^{-11}$	$8{\cdot}96 \times 10^{-11}$	$9{\cdot}08 \times 10^{-11}$	$8{\cdot}77 \times 10^{-11}/8{\cdot}85 \times 10^{-11}$	0·9998

[a] Windig and Antalek result.[13]
[b] Lower/upper limits of 99% confidence.
[c] Correlation coefficient.

Table 2. Self-diffusion coefficients and linearity of gelatin/di-(C6-Glu) mixture

	Self-diffusion coefficient ($m^2\ s^{-1}$)				
	Slope	Eigenvalue	Direct	PMF	r
Gelatin	$1{\cdot}23 \times 10^{-11}$	$1{\cdot}23 \times 10^{-11}$	$1{\cdot}27 \times 10^{-11}$	$1{\cdot}20 \times 10^{-11}/1{\cdot}24 \times 10^{-11}$	0·9999
Di-(C6-Glu)	$2{\cdot}32 \times 10^{-10}$	$2{\cdot}35 \times 10^{-10}$	NA	$2{\cdot}13 \times 10^{-10}/2{\cdot}24 \times 10^{-10}$	0·9998

See footnotes to Table 1.

Table 3. Self-diffusion coefficients and linearity of 2-aminobenzothiozole/2-chloropropionic acid/DMSO/water mixture

	Self-diffusion coefficient ($m^2\ s^{-1}$)				
	Slope	Eigenvalue	Direct	PMF	r
DMSO/water	$7{\cdot}27 \times 10^{-10}$	$7{\cdot}32 \times 10^{-10}$	$4{\cdot}92 \times 10^{-10}$	$7{\cdot}078 \times 10^{-10}/7{\cdot}26 \times 10^{-10}$	0·9999
2-Aminobenzothiozole	$3{\cdot}10 \times 10^{-10}$	$3{\cdot}10 \times 10^{-10}$	$3{\cdot}00 \times 10^{-10}$	$2{\cdot}81 \times 10^{-10}/2{\cdot}83 \times 10^{-10}$	1·0000
2-Chloropropionic acid	$3{\cdot}91 \times 10^{-10}$	$3{\cdot}91 \times 10^{-10}$	$3{\cdot}54 \times 10^{-10}$	$3{\cdot}889 \times 10^{-10}/3{\cdot}90 \times 10^{-10}$	1·0000

See footnotes to Table 1.

The correlation coefficient is 0·995. The agreement between the extracted and reference spectrum of TX-100 is also good, with a correlation coefficient of 0·930. These values are essentially identical to those of Windig and Antalek.

The extracted spectra of gelatin and di-(C6-Glu) are plotted in Figure 2 together with their reference spectra. Good agreement was obtained between the calculated and reference spectrum of gelatin, with a correlation coefficient of 0·993. It is clear from the figure that the main spectral features of di-(C6-Glu) were also extracted. However, a lower correlation coefficient of 0·794 was obtained. There is a serious spectral overlap between the spectra of gelatin and di-(C6-Glu), resulting in some residual rotational freedom. The results presented here are the direct output of the algorithm without further adjustment. The correlation coefficients for these two components when derived by GRAM are 0·993 and 0·800, respectively,[13] showing a similar problem with fully resolving the di-(C6-Glu) spectrum.

Including water and the solvent, DMSO, mixture 3 actually contains four components. However, similar to the GRAM results, only three significant components could be obtained. The extra factor in a four-factor model can be characterized as noise (see Figure 4). Shown in Figure 3 it appears that a combined component of DMSO and water was derived. The two resolved components represent 2-aminobenzothiozole and 2-chloropropionic acid respectively. Since reference spectra were not provided for these components, no correlation coefficients between the calculated and reference spectra were computed.

By regressing the natural logarithm of the factor scores representing the compound concentrations in each mixture against time, excellent linear relationships were obtained. The correlation coefficients of linear regressions are listed in Tables 1–3. The 99% confidence intervals for the slopes were also calculated and then the corresponding confidence ranges of the self-diffusion coefficients of the components were calculated using equation (6). The values are listed in Tables 1–3 along with those of Windig and Antalek.[13] There are no significant differences between the two sets of results.

CONCLUSIONS

The results show that PMF can perform curve resolution. The spectra, the concentration profiles and the subsequent estimates of self-diffusion coefficients of the resolved components are consistent with those obtained by GRAM. The results of this paper together with those from our previous study[11,12] suggest that PMF can be useful for mixture resolution problems.

This investigation shows the possibility of applying PMF to analytical curve resolution, although in this case we have not made use of the features that a least squares method permits and which cannot be applied in an eigenvector analysis. Non-negativity constraints are a useful feature of PMF, since most physically measurable quantities are theoretically non-negative. Such constraints are generally very helpful in environmental applications of PMF.[7–10] However, the results of this investigation showed that bias was introduced when non-negativity was active. Small-amplitude negative values derive from the random noise of the instrument. Forcing all the negative values positive will change the data structure and thus introduce bias. This bias produced active non-negativity constraint results that were slightly worse than those obtained when the constraint was off.

Considering the speed of the processing, PMF is slower than GRAM, since PMF is an iterative fitting algorithm. However, real data contain noise or distortions from the ideal error-free model, and direct algorithms may not always find the optimum solution.[12] Moreover, direct algorithms such as GRAM and DTD are unable to include data point weighting or constraints such as non-negativity, unimodality, etc. which in many cases may be useful in finding the optimum solution.

ACKNOWLEDGEMENTS

The work at Clarkson University was supported by the National Science Foundation under grant ATM 9523731. P. Paatero acknowledges the financial support from the Vilho, Yrjö and Kalle Väisälä Foundation.

REFERENCES

1. F. C. Sanchez, B. van den Bogaert, S. C. Rutan and D. L. Massart, *Chemometrics Intell. Lab. Syst.* **34**, 139–171 (1996).
2. R. A. Harshman, *UCLA Working Papers Phonet.* **16**, 1–84 (1970).
3. J. B. Kruskal, *Linear Algebra Appl.* **18**, 95–138 (1977).
4. K. S. Booksch and B. R. Kowalski, *J. Chemometrics*, **8**, 287–292 (1994).
5. P. K. Hopke, *Receptor Modeling for Air Quality Management*, Elsevier, New York (1991).
6. P. Paatero and U. Tapper, *Environmetrics*, **5**, 111–126 (1994).
7. S. Juntto and P. Paatero, *Environmetrics*, **5**, 127–144 (1994).
8. P. Anttila, P. Paatero, U. Tapper and O. Jarvinen, *Atmos. Environ.* **29**, 1705–1718 (1995).
9. A. V. Polissar, P. K. Hopke, W. C. Malm and J. F. Sisler, *Atmos. Environ.* **30**, 1147–1157 (1996).
10. Y. L. Xie, P. K. Hopke, P. Paatero, L. A. Barrie and S. M. Li, *J. Atmos. Sci.* in press.
11. P. Paatero, *Chemometrics Intell. Lab. Syst.* **37**, 23–35 (1997).
12. P. K. Hopke, P. Paatero, H. Jia, R. T. Ross and R. A. Harshman, *Chemometrics Intell. Lab. Syst.* in press.
13. W. Windig and B. Antalek, *Chemometrics Intell. Lab. Syst.* **37**, 241–254 (1997).
14. E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **4**, 29–45 (1990).

