ENVIRONMETRICS, VOL. 5, 111–126 (1994)

POSITIVE MATRIX FACTORIZATION: A NON-NEGATIVE FACTOR MODEL WITH OPTIMAL UTILIZATION OF ERROR ESTIMATES OF DATA VALUES*

PENTTI PAATERO AND UNTO TAPPER†

University of Helsinki, Department of Physics, Siltavuorenpenger 20 D, SF-00170 Helsinki, Finland

SUMMARY

A new variant 'PMF' of factor analysis is described. It is assumed that $\mathbf{X}$ is a matrix of observed data and $\boldsymbol{\sigma}$ is the known matrix of standard deviations of elements of $\mathbf{X}$. Both $\mathbf{X}$ and $\boldsymbol{\sigma}$ are of dimensions $n \times m$. The method solves the bilinear matrix problem $\mathbf{X} = \mathbf{GF} + \mathbf{E}$ where $\mathbf{G}$ is the unknown left hand factor matrix (scores) of dimensions $n \times p$, $\mathbf{F}$ is the unknown right hand factor matrix (loadings) of dimensions $p \times m$, and E is the matrix of residuals. The problem is solved in the weighted least squares sense: $\mathbf{G}$ and $\mathbf{F}$ are determined so that the Frobenius norm of $\mathbf{E}$ divided (element-by-element) by $\boldsymbol{\sigma}$ is minimized. Furthermore, the solution is constrained so that all the elements of $\mathbf{G}$ and $\mathbf{F}$ are required to be non-negative. It is shown that the solutions by PMF are usually different from any solutions produced by the customary factor analysis (FA, i.e. principal component analysis (PCA) followed by rotations). Usually PMF produces a better fit to the data than FA. Also, the result of PF is guaranteed to be non-negative, while the result of FA often cannot be rotated so that all negative entries would be eliminated. Different possible application areas of the new method are briefly discussed. In environmental data, the error estimates of data can be widely varying and non-negativity is often an essential feature of the underlying models. Thus it is concluded that PMF is better suited than FA or PCA in many environmental applications. Examples of successful applications of PMF are shown in companion papers.

KEY WORDS Factor analysis Principal component analysis Weighted least squares Alternating regression Error estimates Scaling Repetitive measurements

1. INTRODUCTION

This work deals with sets of similar measurements. One measured result is a sequence of numbers. It could be a spectrum, or a set of concentrations for different chemical compounds, or a record of a concentration or intensity as a function of time. The set of similar measured results is a set of sequences of numbers, i.e. a matrix of numbers. The rows or columns of the matrix are assumed to be similar in such a way that they may consist of common basis components. It should be possible to form meaningful approximations of the rows or columns by means of these basis components. Furthermore, it is assumed that this basis is unknown, as well as the coefficients used for representing the matrix with this basis.

It turns out that the problem is symmetric with respect to the basis and the coefficients, this is a

* Fourth International Conference on Statistical methods for the Environmental Sciences 'Environmetrics', August 17 to 21, 1992, Espoo, Finland

† Present address: Technical Research Centre of Finland (VTT), Aerosol Technology Group, P.O. Box 206, SF-02151 Espoo, Finland

CCC 1180–4009/94/020111–16

Received 19 April 1993
Revised 25 January 1994

'bilinear model'. The original matrix is approximated by a product of two smaller unknown matrices. The model is also called a 'factor model'. Principal component analysis (PCA) and factor analysis (FA) are known methods for solving a bilinear model. These methods have been very successful in such fields as social sciences and psychometry. However, in physical sciences they have found only limited application.

The method being introduced utilizes error estimates of elements of the measured data matrix and implements strict non-negativity constraints for the factors, i.e. for the basis and for the coefficients used for approximating the matrix. Thus the model is more quantitative and better suited for physics, chemistry, and environmental sciences than the customary PCA-based models. Successful applications of the method to a few aerosol problems have been published earlier by Paatero *et al.*[1]

The connection of error estimates for the matrix and scalings used as a preliminary step of PCA was investigated in an earlier paper by Paatero and Tapper[2]. It was shown that a correctly weighted least squares formulation of the factor problem leads to an optimal fitting of the data matrix. Also it was demonstrated that the customary R-mode and Q-mode forms of factor analysis are equivalent to solving weighted least squares problems with unrealistic error estimates for the observed data matrix. Some of those results are summarized in this work.

Application of the method to practical environmental data requires that error estimates for the data be chosen judiciously so that the estimates reflect the quality and reliability of different data points. For a typical environmental problem, with outliers and other problematic data points, this is reported in detail by Tapper.[3] For 'clean' data, e.g. from nuclear spectroscopy, the usual error estimates (square root of count) can be used as such without any problems.

The non-negativity constraint is natural in many physical problems; there cannot be a negative amount of a basic constituent in any sample, nor can the composition of any basic constituent contain a negative percentage of any element. The factor model should obey these constraints so that there are no negative entries in the factor matrices. The present standard approach (PCA), however, can produce negative values for almost all factors. The problem of transforming ('rotating') factors in order to eliminate the negative entries has been discussed by several authors, see, for example, Henry[4] and Shen and Israël.[5] The solutions offered so far do not seem satisfactory; there usually remain some negative values in the factors. The reason will become apparent later on.

The technique of alternating regression (AR) is rather similar to the present one. It has found wide use in chemometrics, typically for analysing results produced by connected instruments, e.g. gas chromatograph–mass spectrometer (GC–MS), see Karjalainen and Karjalainen.[6,7] In AR, the error estimates for data values have not been explicitly used. Instead, various strong *a priori* constraints are routinely applied, such as the unimodal shape of the 'peak shape' factors. Also, different algorithms have been found more suitable for AR than for PMF. The data which is available in typical cases of AR has much better information content than a typical set of environmental data. Thus it is not as essential to extract the last bit of information in AR as it is in a typical application of PMF. The mathematical/statistical analysis of the AR approach is hampered by the fact that the AR approach is not defined as a minimization problem. Instead, it is defined by specifying the algorithm used for computing the model. Thus an exact comparison of AR with the present approach is quite difficult.

2. TERMINOLOGY AND NOTATION

The term 'factor analysis' (FA) is ambiguous. In chemometrics and aerosol physics, FA means principal component analysis (PCA): singular value decomposition (SVD), selection of

dimension, and rotations. Statisticians often remark that this is not FA at all according to their definition of FA. In statistics, FA means investigation of correlations of random variables. This leads to a non-linear computation which cannot be done with SVD. This statistical FA is sometimes called 'orthodox FA' or 'non-linear FA' by physicists or chemists. It is seldom useful in physics or chemistry. The subject of the present work is similar to the chemical variant of FA. We are working with mathematical models which cannot be solved with SVD. Thus we would rather not call this PCA. Also, it is good to avoid the ambiguous term 'FA'. We suggest that our approach should be called 'positive matrix factorization' or PMF.

2.1. Variables and observations

In many applications, different columns of $\mathbf{X}$ contain measurements made of different quantities, e.g. concentrations of different compounds or elements. Then each column is said to correspond to one *variable*, the words 'column' and 'variable' are used interchangeably. Then also different rows may be called *observations*. Because of the symmetry of the problem, rows could also correspond to variables and columns to observations. On the other hand, there are also applications where the notion of variables and observations is quite meaningless. Measuring the intensity of an optical spectrum as a function of wave length and time is such an example. In such an application, one should interpret the word 'variable' as 'row or column of matrix $\mathbf{X}$'.

$||\mathbf{B}||_F$ The Frobenius norm of $\mathbf{B} = \sqrt{(\sum_i \sum_j \mathbf{B}_{ij}^2)}$ arg min. The expression arg $\min_\theta f(\theta)$ denotes that value of θ which minimizes $f(\theta)$

PCA Principal component analysis

FA Factor analysis, especially 'orthodox factor analysis'

LS Least squares

SVD SVD of $\mathbf{X}$ means singular value decomposition: $\mathbf{X} = \mathbf{USV}'$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{S}$ is diagonal with non-increasing entries.
The prime denotes the transpose of a matrix. The matrices are denoted by bold face letters, mostly in upper case
The tilde denotes part of a matrix which is selected according to the dimension of the factor model: $\tilde{\mathbf{U}}$ is a matrix which consists of the first p columns of $\mathbf{U}$

PMF 'Positive matrix factorization', finding weighted LS solutions $\mathbf{G}$ and $\mathbf{F}$ for the equation $\mathbf{X} = \mathbf{GF} + \mathbf{E}$, where elements of the unknown factor matrices $\mathbf{G}$ and $\mathbf{F}$ are required to be non-negative

$\mathbf{X}$ The observed matrix of dimensions $n \times m$ of elements $\mathbf{X}_{ij}$

i The index i denotes rows of $\mathbf{X}$ and/or $\mathbf{A}$, $i = 1, \ldots, n$

j The index j denotes columns of $\mathbf{X}$ and/or $\mathbf{A}$, $j = 1, \ldots, m$

$\mathbf{A}$ A scaled version of the original $\mathbf{X}$. Decompositions of $\mathbf{A}$ are sometimes written in lowercase letters, $\mathbf{A} = \mathbf{usv}' + \mathbf{e}$

σ The matrix of error estimates (standard deviations) of elements $\mathbf{X}_{ij}$

$\mathbf{D}$ and $\tilde{\mathbf{D}}$ Diagonal matrices used for computing scaled versions of $\mathbf{X}$, e.g. $\mathbf{A}_L = \mathbf{DX}$ or $\mathbf{A}_R = \mathbf{X}\tilde{\mathbf{D}}$.

3. SETTING UP THE FACTORIZATION PROBLEM

Traditionally the factorization or PCA has been based on the correlation matrix. Later, the singular value decomposition (SVD) has also been used. The first alternative requires that the data matrix be first centred; this results in a loss of information about the origin of the scale of variables. On the other hand, physical and chemical models deal with quantitative scales having a

well-defined origin or zero point. For example, the entries in matrices may represent observed mass values. Then zero means 'no mass' and negative values cannot occur. (This is in contrast to many scales in social sciences, say, where zero is just a matter of convention.) Thus centring would cause loss of information and it would also prevent the use of non-negativity constraints. Hence we do not centre any variables. We consider that approaching the factorization problem through covariance matrix is inappropriate in the physical sciences! We shall not be satisfied in finding some correlations, we wish to form a quantitative model of what was observed.

The factor model is thus: $\mathbf{X} \approx \mathbf{GF}$, where $\mathbf{X}$, $\mathbf{G}$ and $\mathbf{F}$ are $n \times m$, $n \times p$, and $p \times m$, respectively. This can be viewed as a sum: $\mathbf{X} \approx \mathbf{X}^1 + \mathbf{X}^2 + \cdots + \mathbf{X}^p$, where the matrices $\mathbf{X}^1, \ldots, X^p$ are of rank one. SVD gives a solution for the factorization problem: if $\mathbf{X} = \mathbf{USV}'$, then the solution of rank p is $\mathbf{X} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}' + \mathbf{E} = \mathbf{GF} + \mathbf{E}$ and it can be shown that this solution has a least squares property: among all approximations of rank p of $\mathbf{X}$, it minimizes $||\mathbf{E}||_F$. We base our definition of the factorization problem on this LS property.

We define the unweighted factorization problem as follows.
Given $\mathbf{X}$, the unweighted factorization of rank p of $\mathbf{X}$ is defined by

$$\{\mathbf{G}, \mathbf{F}\} = \arg \min_{\mathbf{G},\mathbf{F}} ||\mathbf{X} - \mathbf{GF}||_{\mathrm{F}} \tag{1}$$

where $\mathbf{G}$ and $\mathbf{F}$ are required to be of previously selected rank p. Because this is defined as an unweighted LS problem, its solution is optimal (i.e. minimum variance) if and only if all the variables $\mathbf{X}_{ij}$ to be fitted are of same accuracy. The unweighted solution can be obtained through SVD.

Remark: this definition is fully symmetric with respect to $\mathbf{G}$ and $\mathbf{F}$; transposition of $\mathbf{X}$ only causes the interchange and transposition of $\mathbf{G}$ and $\mathbf{F}$, otherwise the problem stays the same and the same approximation will be obtained for $\mathbf{X}$. This symmetry will be valid for all of our discussions of PMF. There is no need to maintain the distinction between 'loadings' and 'scores'.

3.1. Different models for randomness

The statistical models underlying orthodox FA and PMF are quite different. In FA, one assumes that each row of the observed matrix is sampled from a multivariate distribution which stays the same from row to row. The covariance matrix of this distribution may be denoted by $\mathbf{\Sigma}$. The matrix includes both the randomness due to the sampling and the randomness due to measurement errors. Thus it is implicitly assumed that measurement errors are of the same magnitude from row to row. In PMF, no sampling process is assumed. Except for the measurement errors, the model is fully deterministic; the unknown factor matrices determine the 'correct' matrix, $\mathbf{X} = \mathbf{GF}$, which cannot be observed. Instead, one observes the matrix $\mathbf{X} = \mathbf{GF} + \mathbf{E}$, where $\mathbf{E}$ is the matrix of random measurement errors. Before performing PMF, one must somehow estimate the size of these errors, i.e. the standard deviations of each element of the matrix $\mathbf{E}$. The standard deviation must be estimated separately for each error term $\mathbf{E}_{ij}$. Together, these standard deviations form a matrix $\boldsymbol{\sigma}$ which is not to be confused with the covariance matrix $\mathbf{\Sigma}$.

4. CONNECTION OF ERROR ESTIMATES AND SCALINGS OF DATA MATRIX

Any weighted LS fit to a model $y_i = y(x_i) = f(x_i, \theta)$ is performed by the minimization

$$\theta = \arg \min_{\theta} \sum_i w_i (y_i - f(x_i, \theta))^2.$$

In order to get a best possible (i.e. minimum variance) solution, the weights must be $w_i = (\text{std.dev}(y_i))^{-2}$.

We assume that std.dev($\mathbf{X}_{ij}$) = $\boldsymbol{\sigma}_{ij}$ are known. The most general scaled form of PCA is defined with the help of diagonal scaling matrices $\mathbf{D}$ and $\bar{\mathbf{D}}$ in the following way: the scaled matrix is $\mathbf{A} = \mathbf{D}\mathbf{X}\bar{\mathbf{D}}$, PCA is based on the SVD of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{D}\mathbf{X}\bar{\mathbf{D}} = \mathbf{usv}'$$

$$= \mathbf{gf} + \mathbf{e}.$$

This scaled solution minimizes $||\mathbf{e}||_F$, the norm of the scaled residual matrix. From this we get an equation for the original unscaled $\mathbf{E}$:

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} = \mathbf{D}^{-1}\mathbf{gf}\bar{\mathbf{D}}^{-1} + \mathbf{D}^{-1}\mathbf{e}\bar{\mathbf{D}}^{-1}$$

$$\mathbf{e} = \mathbf{DE}\bar{\mathbf{D}}.$$

As expressed with the unscaled residual matrix $\mathbf{E}$, the solution of the scaled problem minimizes

$$||\mathbf{e}||_F = ||\mathbf{DE}\bar{\mathbf{D}}||_F = \left(\sum\sum \mathbf{D}_{ii}^2\mathbf{E}_{ij}^2\bar{\mathbf{D}}_{jj}^2\right)^{1/2}. \tag{2}$$

The weights in this LS minimization are $w_{ij} = \mathbf{D}_{ii}^2\bar{\mathbf{D}}_{jj}^2$. The optimality condition for scaled PCA requires that each weight equals the inverse of the square of the standard deviation of $\mathbf{E}_{ij}$, or that

$$\boldsymbol{\sigma}_{ij} = \mathbf{D}_{ii}^{-1}\bar{\mathbf{D}}_{jj}^{-1}. \tag{3}$$

The right side of this equation is always of rank one. Thus the condition can only be fulfilled if $\boldsymbol{\sigma}$ is of rank one too. The solution of the factorization problem by scaled SVD can only be optimal if the standard deviation matrix is of rank one! Example: the column norm scaling, where $\mathbf{D} = I$ and $\bar{\mathbf{D}}_{jj} = (\sum_h \mathbf{X}_{hj}^2)^{-1/2}$, is optimal if and only if $\boldsymbol{\sigma}_{ij} = \text{Const}\,(\sum_h \mathbf{X}_{hj}^2)^{1/2}$.

In real experimental situations, the matrix $\boldsymbol{\sigma}$ is known (or estimated). If it is of rank one, then the scaling matrices $\mathbf{D}$ and $\bar{\mathbf{D}}$ can and should be determined so that the optimality condition be fulfilled. If, however, rank($\boldsymbol{\sigma}$) > 1, then the optimality condition cannot be fulfilled and SVD cannot be an optimal LS method. A near-optimal method can be obtained by approximating $\boldsymbol{\sigma}$ by a matrix $\bar{\boldsymbol{\sigma}}$ of rank one and performing a scaling of $\mathbf{X}$ from left and right so that the scaling is optimal with respect to this approximate standard deviation matrix $\bar{\boldsymbol{\sigma}}$. This method can be significantly better than the customary column norm scaling, as shown by Paatero and Tapper.[2]

5. POSITIVITY CONSTRAINTS AND 'ROTATIONS' WITH PCA

It is well known that any non-singular matrix $\mathbf{T}$ defines a rotation of the solution by

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} = \mathbf{GTT}^{-1}\mathbf{F} + \mathbf{E} = \bar{\mathbf{G}}\bar{\mathbf{F}} + \mathbf{E}, \tag{4}$$

where the new rotated factors are $\bar{\mathbf{G}} = \mathbf{GT}$ and $\bar{\mathbf{F}} = \mathbf{T}^{-1}\mathbf{F}$. A rotation does not affect the residual matrix $\mathbf{E}$.

As an example of rotations, we wish to introduce the following important pair of matrices:

$$\mathbf{T} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}, \quad \mathbf{T}^{-1} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}. \tag{5}$$

The matrices in equation (5) represent addition of one column of $\mathbf{G}$, multiplied by a scalar factor a, to another column, and subtraction of one row of $\mathbf{F}$, multiplied by the same scalar factor a,

from another row. All rotations can be represented as sequences of these elemental rotations. Thus these elemental rotations form a basis for all possible rotations. If none of these elemental rotations is possible in a given situation, then we might conclude that all rotations are impossible.

These matrices are not rotations in the proper mathematical sense. However, in the jargon of FA or PCA, all linear transformations are called 'rotations' and we are following this unfortunate practice.

If the matrix **X** has been centred it contains both positive and negative elements and it cannot be approximated by non-negative factor matrices. For this reason, we recommend that **X** should not be centred; then it usually consists of positive elements and one may hope to achieve a non-negative factorization.

Rotations can be applied for eliminating negative elements from factors **G** and **F**. However, increasing the values on one side (G/F) simultaneously decreases the values on the other side (F/G) because one of the coefficients a and $-a$ in the matrices T and T^{-1} in equation (5) must be negative. Thus there is the risk of producing new negative elements. This problem is discussed for example, by Henry[4] and Shen and Israël;[5] they suggest sequences of linear programming or sequences of rotations for solving the problem.

Definition: a factorization $\mathbf{X} = \mathbf{GF} + \mathbf{E}$ of selected rank r is called p-rotatable (p means 'positive-') if it can be transformed ('rotated') to a different factorization $\mathbf{X} = \mathbf{GTT}^{-1}\mathbf{F} + \mathbf{E}$ so that all the elements of the new factors **GT** and $\mathbf{T}^{-1}\mathbf{F}$ be non-negative and **T** be not diagonal.

Remark: there exist matrices consisting of non-negative entries such that the SVD-derived factorization of the matrix is not p-rotatable.

Example: the following optimal factorization of rank two of the matrix **X** is not p-rotatable. (The matrices are shown rounded to two decimal places.)

$$\mathbf{X} = \begin{vmatrix} 100 & 10 & 0 \\ 10 & 0{\cdot}8 & 0 \\ 0 & 1 & 0 \end{vmatrix} = \begin{vmatrix} 1{\cdot}00 & 0{\cdot}01 \\ 0{\cdot}10 & -0{\cdot}10 \\ 0{\cdot}00 & 1{\cdot}00 \end{vmatrix} \times \begin{vmatrix} 100{\cdot}00 & 9{\cdot}98 & 0{\cdot}00 \\ -0{\cdot}10 & 1{\cdot}00 & 0{\cdot}99 \end{vmatrix} + \begin{vmatrix} 0{\cdot}00 & 0{\cdot}01 & -0{\cdot}01 \\ 0{\cdot}01 & -0{\cdot}10 & 0{\cdot}10 \\ 0{\cdot}00 & -0{\cdot}01 & 0{\cdot}01 \end{vmatrix}.$$

The element $\mathbf{F}_{21} = -0{\cdot}1$ could be annihilated by adding the first row of **F** to the second, multiplied by 0·001. This would make a smaller negative entry $\mathbf{G}_{31} = -0{\cdot}001$ which could not be annihilated. An attempt to annihilate the element $\mathbf{G}_{22} = -0{\cdot}1$ by adding the first column of **G** to the second would create a larger negative entry $\mathbf{F}_{13} = -0{\cdot}99$. Again, this entry could not be annihilated without reintroducing $\mathbf{G}_{22} < 0$. Thus is not possible to annihilate any of the negative entries by adding rows or columns to each other without producing new negative entries.

There are published algorithms for rotating the solutions so that all factors should become non-negative. These algorithms start with the solution given by SVD and then apply complicated schemes of rotations in order to eliminate the negative entries. The existence of matrices which are not p-rotatable means that these algorithms cannot always succeed in achieving a factorization with non-negative factors, see Shen and Israël.[5]

6. POSITIVE MATRIX FACTORIZATION

6.1. Definition of the weighted factorization problem with non-negativity constraints

Given **X** and $\boldsymbol{\sigma}$ and the selected rank p, we define positive matrix factorization of **X** as follows:

$$\mathbf{X} = \mathbf{GF} + \mathbf{E}, \quad \mathbf{G}: n \times p, \quad \mathbf{F}: p \times m, \tag{6}$$

$$\mathbf{G}_{ik} \geqslant 0, \quad \mathbf{F}_{kj} \geqslant 0, \tag{7}$$

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{E}_{ij}^2 / \sigma_{ij}^2, \tag{8}$$

$$\{\mathbf{G}, \mathbf{F}\} = \arg \min_{\mathbf{G},\mathbf{F}} Q. \tag{9}$$

This is a difficult task because it has two different non-linearities: inequalities and products of unknowns. Special programs have been written by the authors for solving this problem. They are useful at least to such dimensions as $n, m, p = 300, 30, 10$. (One of the dimensions could be larger if another is smaller.) Small problems (10, 5, 3 say) could presumably also be solved by standard numerical software.

7. ALGORITHMS FOR PMF

The algorithms are still being developed further and new principles are being tested. Thus the details of the current algorithms are omitted; they will be obsolete in a few months. The main ideas are given in the following.

The original basic algorithm is alternating regression (AR). Each of the factors **G** and **F** is solved alternately while keeping the other factor constant. In the customary PCA, the final convergence of AR depends on the singular values of **X**: if $\mathbf{S}_{pp} \gg \mathbf{S}_{p+1,p+1}$, then the AR algorithm converges rapidly. Satisfactory use of AR is reported by Karjalainen;[8] in the analysis of chromatograph–mass spectrometer data, AR typically converges in five iteration steps. Our experience indicates that in PMF the final convergence of AR is often hopelessly slow (needing up to thousands of steps) if the factors are far from being orthogonal. This difference is due to two aspects: the factors in typical AR problems are wide overlapping peaks on the wavelength side but quite narrow lines on the mass side, whereas our experience comes from cases where the factors are wide smooth shapes on *both* sides. Also, the constraint of unimodality or other strong constraints in AR contribute towards the fast convergence in AR.

The AR algorithm can be significantly improved by performing extended (**G**, **F**) steps where both **G** and **F** are changing simultaneously. The iteration consists now of the repetition of the following three basic steps: start from $\mathbf{G} = \mathbf{G}_0$, $\mathbf{F} = \mathbf{F}_0$, solve for $\mathbf{G} = \mathbf{G}_0 + \Delta\mathbf{G}$ while keeping $\mathbf{F} = \mathbf{F}_0$ constant; then solve for $\mathbf{F} = \mathbf{F}_0 + \Delta\mathbf{F}$ while keeping $\mathbf{G} = \mathbf{G}_0 + \Delta\mathbf{G}$ constant; and third, solve for the extension coefficient α in

$$(\mathbf{G}_0 + \alpha\Delta\mathbf{G})(\mathbf{F}_0 + \alpha\Delta\mathbf{F}) = \mathbf{X} + \mathbf{E},$$

where $\Delta\mathbf{G}$ and $\Delta\mathbf{F}$ are as determined from the first and second steps. Towards the end of the iteration, α gets quite large values, thus making a significant improvement over the basic AR algorithm.

This iteration has been our work horse. It is fast but the program is tricky. It is difficult to incorporate automatic rotation selection in this algorithm. It is possible to handle extremely large matrices with this algorithm, because only a small number of unknowns need to be solved simultaneously. Typically the convergence of this iteration needs 30 to 100 steps.

The most elegant algorithm consists of simultaneous solutions of $\Delta\mathbf{G}$ and $\Delta\mathbf{F}$ by minimization of $||\mathbf{E}||_F$ in the equation

$$(\mathbf{G}_0 + \Delta\mathbf{G})(\mathbf{F}_0 + \Delta\mathbf{F}) = \mathbf{X} + \mathbf{E},$$

where the second-order term $\Delta\mathbf{G}\Delta\mathbf{F}$ can be ignored. This algorithm converges faster, in 15 to 30 steps. It requires a large amount of memory for a matrix of size $(n + m) \times m \times p^2$. Thus available memory limits the size of problems to be solved. The algorithm is nice and clean and various

rotation selections may be incorporated in it. It is also possible to estimate the variance of factor elements. This algorithm is ready for distribution to interested colleagues.

It seems possible to combine the good properties of the last two algorithms. This is under development. In all of the algorithms the negative entries must be dealt with somehow. In the original AR algorithm, the negative values are simply truncated to zero. We have used a penalty function which is proportional to the square of the negative value. The penalty coefficients are adjusted dynamically; if a component gets a positive value, the corresponding penalty coefficient is decreased towards a very small minimum value. If a component gets a negative value, the corresponding coefficient is increased towards a fixed maximum value. This algorithm is simple and reliable, provided that the penalty coefficients are not changed too much during one iteration step. For usual non-negative least squares, the algorithm may not be very efficient. However, in factorization the LS problem is not yet finally defined during the early stages of the iteration. Thus it does not matter that the penalty values are not yet ready during those early steps.

In standard PCA the function Q has one global minimum value and local minima cannot occur. In PMF, however, the function Q may possess local minima. In different cases this may be caused by the individual weighting of the elements $\mathbf{X}_{ij}$, or by the non-negativity constraints, or by the combined action of both. All of the algorithms for PMF can find a minimum of $Q(\mathbf{G}, \mathbf{F})$ but they do not indicate whether the minimum is a local or a global one. Different sets of pseudorandom numbers can be used as starting points $(\mathbf{G}_0, \mathbf{F}_0)$ in order to identify the global minimum of Q among the local minima. Our experience indicates that local minima do sometimes occur in PMF but in most cases there is only a single global minimum. According to Karjalainen,[8] in AR wrong local minima are never observed in practice. When discussing the properties of PMF results, we shall tacitly assume that a global minimum has always been found.

8. THE ROLE OF THE STANDARD DEVIATION MATRIX $\boldsymbol{\sigma}$

The standard deviation matrix $\boldsymbol{\sigma}$ is an important link between the factor model and the physical reality. Various kinds of information may be communicated to the model by means of setting the standard deviations accordingly. In explaining these techniques, we use the index k to denote one variable and the corresponding column of $\mathbf{X}$. The index i denotes all values $1, \ldots, n$, the index l denotes a single value of i.

Consider the situation where the values of the kth variable are smaller than the noise level of the measuring system or model. In customary PCA this may lead to the formation of a noise factor, consisting mainly of the kth variable. It is the task of the user to recognize such a non-informative factor. When using PMF, the standard deviations would show the situation as $\mathbf{X}_{ik} < \boldsymbol{\sigma}_{ik}$. This would mean that the kth variable would contribute only a little to the Q function defined in equation (8). The values $\mathbf{X}_{ik}$ practically would not influence the formation of factors. The model would ignore the presence of such a noise variable without any help from the user.

In environmental work the distribution of observed values is skewed; very large values occur with non-zero probability. Sometimes these values may be due to errors, sometimes they are correct. In both cases, they disturb non-robust methods such as PCA and FA. As 'outliers', they must be removed from the data matrix, thus causing loss of information. Within the framework of PMF, an especially large observed value $\mathbf{X}_{lk}$ may be handled in the following way: $\mathbf{X}_{lk}$ is given a standard deviation which is a fixed fraction of the observed value, $\boldsymbol{\sigma}_{lk} = 0{\cdot}1\mathbf{X}_{lk}$, say. This will prevent the large value from influencing the factor formation too much. It is not necessary to work with the logarithms of the data unless the original model is multiplicative. It was demonstrated by Paatero and Tapper that unfounded use of logarithms may cause the appearance of ghost factors.[2,9]

Missing data points cause a similar problem as outliers. In traditional PCA, the safest technique is to discard all such rows of X where there are missing values. This may lead to intolerable loss of data. In PMF, a missing data point X_{lk} is handled as follows: the mean value of the kth column $\mathbf{X}_{ik}$ is inserted into $\mathbf{X}_{lk}$ and the standard deviation of the kth column into $\boldsymbol{\sigma}_{lk}$. This essentially tells the model that nothing is known about the value except that it must be similar to the other values of the variable. Especially, it cannot be extremely large.

Handling of outliers can be continued *a posteriori*, after the fit. After the PMF algorithm has converged, the program may calculate the residuals and identify the points of bad fit, i.e. the points where the residual is much larger than the standard deviation, $|\mathbf{E}_{lk}| \gg \boldsymbol{\sigma}_{lk}$. If these are considered to be outliers, their harmful effect can be limited by increasing their standard deviations and rerunning the PMF algorithm with these new values. This approach is safer than the outright rejection of outliers; because there is no good/bad decision to make, moderately bad points get a moderate decrease of significance but they are not rejected altogether. This avoids the risk of randomly rejecting perfectly good points which just happen to get a worse fit.

Handling of outliers can be started *a priori*, before seeing the data values at all. Sometimes one knows that certain observations are more likely to contain large errors (for example, in precipitation studies, measured concentrations of low-volume samples are unreliable). Before starting PMF, one can increase the standard deviations of such unreliable values by a fixed factor, e.g. by a factor of 2.

What errors should be included in standard deviations? All those errors which cause $\mathbf{X}$ to deviate from a best-fit matrix of rank p. Thus sampling errors of random nature (e.g. caused by biological transformations in the sampling container or by local contamination from peculiar sources) should be included. Estimate of a possible constant laboratory bias should be excluded.

8.1. Subjective emphasis

With few variables (8, say) only a small number of factors (3, say) can be extracted. Some of the variables can be more important than others, e.g. the element Na might be almost irrelevant when air pollutants are studied. One can subjectively increase the values $\boldsymbol{\sigma}_{ik}$ for such a variable. This way the available factors are not 'wasted' in explaining the behaviour of the suppressed irrelevant variable – this, of course, is a problematic question.

In selectively increasing the standard deviations, one might be afraid of losing the objectiveness of the analysis. This is a real risk. However, it must be weighed against what else is available; if a large value is rejected as a suspected outlier, then it in fact gets zero weight, corresponding to an infinite standard deviation. Giving it a large but finite standard deviation must necessarily be less risky than rejecting it. Both techniques tend to introduce a bias towards small values. However, if the large values themselves were originated by a process which is outside of the model being studied, then they constitute a bias towards large values and deleting their influence simply removes this bias. Also, increasing the standard deviations of an irrelevant variable is less drastic than omitting the variable altogether from the analysis or not measuring it at all.

8.2. Unique factors

Sometimes one accidentally specifies too small values $\boldsymbol{\sigma}_{ik}$ for some variable k. This causes the noise of that variable to appear significant. In PMF such a variable is then likely to form a unique factor, i.e. a factor where on the row of matrix $\mathbf{F}$ only one variable has a significantly non-zero value (the same is true in PCA). To complicate the situation, the variable then gets a very good fit, perhaps misleading the unwary to infer that the specified standard deviations were too large! In

precipitation studies it was observed that the concentration of the element K tended to form a unique factor.

All unique factors are not necessarily due to noise, however. It may happen that one significant column of **X** is so different from others that it deserves a factor of its own, a unique factor. Such a result is meaningful in many fields. We have encountered examples in precipitation data and in X-ray diffraction spectra measured as functions of temperature. The orthodox FA would fail here, as it is only modelling *common* factors, i.e. factors which reflect correlation of columns of **X**.

9. COMPARISON OF PMF AND PCA

Selecting the dimensions or rank of the model or 'number of factors' p is a similar problem in both models; the decisive question is which number is most useful. The answer is only obtained by trying different dimensions. In PCA, the solutions form a hierarchy; a higher dimension contains all the factors of the lower dimensions. However, finding the most useful rotations in PCA is a different task for every dimension. In PMF the factors are not orthogonal and there is no hierarchy. All factors may change when the dimension is increased. In the following we assume that the dimension has already been selected (the same value for PMF and for PCA). Also we assume that the standard deviations σ_{ij} given to PMF are correct.

The factors generated by PCA are orthogonal. If general ('oblique') rotations are applied, then the rotated factors are not orthogonal any more. The factors generated by PMF are not orthogonal at all. Orthogonality may be a desirable property, e.g. in psychometry. In the physical sciences, however, the desired factors are usually some physical distributions (e.g. the time behaviour of a quantity or a concentration profile of a pollution source). Such distributions are practically never orthogonal, thus the orthogonal property of PCA is useless. On the contrary, many physical distributions should be non-negative which precludes orthogonality. PMF produces non-negative distributions (i.e. factors) by definition. When using PCA, one may attempt to perform rotations in order to fulfil the non-negativity constraints as far as possible.

One drawback of PMF is that the programs for PMF are much slower. On the other hand, they may be set up to run fully automatically without human guidance in the majority of cases where the non-negativity criteria lead to a good rotation without human help.

The detailed comparisons depend on the rank of standard deviation matrix $\boldsymbol{\sigma}$ and whether the factorization by SVD is p-rotatable or not. The different cases are mentioned separately in the following.

9.1. The matrix $\boldsymbol{\sigma}$ is of rank one and SVD of X is p-rotatable

Here the matrix **X** can be scaled correctly for computing the SVD. Thus the factorization by SVD is also optimal (i.e. minimum variance). Of course, the factorization by PMF is always optimal because it always utilizes the correct standard deviations. When the solution given by SVD is p-rotated, then it becomes a solution of the PMF task, because both have the same residuals.

The two methods are equivalent and $Q(\text{PMF}) = Q(\text{PCA})$. However, PMF produces a desired non-negative solution directly, whereas the solution by PCA must be rotated in order to obtain a non-negative solution.

9.2. The matrix $\boldsymbol{\sigma}$ is of rank one but SVD of X is not p-rotatable

In this case it is impossible to rotate the SVD-derived factorization so that both **G** and **F** would be strictly non-negative. On the other hand, PMF will produce the desired solution. The solution by

PMF is a worse approximation, $Q(\text{PMF}) > Q(\text{PCA})$, because PMF does not cheat by using negative entries in factors as the SVD does. PMF solves the problem, PCA does not.

9.3. Rank(σ) > 1

Correct scaling is not possible with SVD. The solution by SVD is only possible so that the true σ is approximated by a matrix of rank one, leading to loss of information when solving the approximate problem. As above, it may also happen that the solution by SVD is not p-rotatable, preventing the solution by PCA. On the other hand, PMF solves the original problem correctly. If the approximate solution obtained with SVD is p-rotatable, then $Q(\text{PMF}) < Q(\text{PCA})$, where both Q values are defined with the true standard deviation. In this case, PMF is producing a better approximation of **X**.

PMF solves the problem. At best, PCA solves an approximate problem, or else it cheats by using negative entries.

In the second and third cases, above, the solutions by PMF may or may not be p-rotatable. If it is not, the factorization result by PMF is unique. This is a valuable property, in contrast to the general ambiguity of the results by PCA or FA.

To sum up, by applying a best possible scaling of rank one, PCA could be made better than what is the customary practice of column norm scaling. However, PMF is better than the best possible PCA, or in some cases equivalent to PCA. The only merit for PCA is that computing a PCA(SVD) is much faster than computing a PMF.

9.4. Objectivity and human guidance

Very often the scientist must interact with the computational process when analysing experimental data. The most important reason for interaction is that the scientist knows which solutions are or are not meaningful in the given context. Preferably this interaction should happen so that it may be objectively described afterwards.

If one searches for a good rotation by trial and error, then the result is subjective and the path to the result cannot be objectively described. Another scientist might arrive at another solution when solving the same problem. Selecting a starting point for the iteration might also be a subjective component in the analysis. In some deconvolution methods the user must make a subjective decision about when a good degree of convergence of the algorithm has been reached.

A better approach is to make the algorithms fully automatic but to include numerical parameters for controlling them. The standard deviations of observed data are an example of such parameters. Selecting the rule for computing the standard deviations of environmental data may be a subjective decision. However, this decision should be reported together with the results and another scientist would be able to replicate the analysis.

Visual checking of outliers is a subjective process and difficult to repeat by another person. On the other hand, the suggested handling of outliers (both *a priori* and *a posteriori*) can be described exactly, even if it contains a subjective component in selecting the actual numerical values for modifying the standard deviations, (see Tapper,[3] and Juntto and Paatero[9]).

One can see that PMF is objective in the sense that its algorithm is fully automatic. PMF does not require the kind of human guidance which would be difficult to describe objectively; the scientist may exert subjective control of PMF in such a way that it may be accurately reported and later replicated at will.

10. UNIQUENESS OF PMF RESULTS

Is the solution p-rotatable? If it is, then one may freely rotate the solution without increasing the Q value and without violating the non-negativity constraints. If such 'free rotations' are possible, then the solution *is not unique*, and by starting the iteration with different random values one might find another solution. If large values are obtained for variance estimates of factor elements, this could perhaps act as an indicator, signalling the possibility of free rotations. Usually, however, a visual inspection is sufficient; if no factor can be subtracted from another without producing negative entries in factor matrices, then the result is unique, i.e. not p-rotatable.

How to favour one solution among a set of p-rotatable solutions? And how to do it in a repeatable objective way? (see PCA where varimax is one technique for selecting among all solutions of rank p.) The minimization task of PMF can be changed to

$$\{\mathbf{G}, \mathbf{F}\} = \arg \min_{\mathbf{G},\mathbf{F}} (Q(\mathbf{G}, \mathbf{F}) + q(\mathbf{G}) + \bar{q}(\mathbf{F})), \tag{10}$$

where q and $\bar{q}$ are application-dependent quadratic functionals, e.g., q can be chosen so that it makes columns of **G** as smooth as possible. With non-trivial q and $\bar{q}$ the degeneracy of a PMF problem usually disappears. Then the solution of PMF is unique.

Another approach for guiding the solution towards a desired rotation is to lock selected factor components to zero. If this increases the Q value significantly, then the component has a non-zero value and the lock must be removed. If, however, the Q value remains the same, then the lock is permissible and can be used for specifying *a priori* information about the rotations.

Different standard deviation matrices can lead to different results by suppressing or enhancing different variables.

The following rule of thumb tells when to expect a unique factorization: if all columns of the correct unknown **G** and all rows of the correct unknown **F** contain a significant number of zeros, then the result by PMF is unique. The Quail Roost II test case is such an example, see Currie *et al.*[10]

If one side (**G**, say) does contain zeros but the other side does not, then the basic PMF result is not unique but an automatic rotation option produces a unique result, maximizing the number of zero entries on the specified side. Such a result is often physically meaningful.

If, however, there are practically no zero entries in any of the correct factors, then there is a neighbourhood of the correct solution where the non-negativity constraints are not active at all. Then the PMF is equally ambiguous as PCA. Many rotation variants can be tried: maximizing the number of zero entries on either the **G** side or the **F** side; finding a minimum norm solution (leading to no zero entries on either side); maximizing the smoothness of some selected factors, etc. It is quite possible that none of these alternatives is physically founded. Their usefulness should be demonstrated in each application separately. In these cases one must be aware that selecting a rotation in PMF really is a subjective decision which must be justified by arguments which are relevant to the application. The same criticism applies to rotations used in PCA; there is really no *a priori* reason why the varimax rotations should produce a physically meaningful solution. It is up to each application to demonstrate the usefulness of any such technique.

11. PRESENTATION OF RESULTS

Either the columns of **G**, or the rows of **F** should be normalized so that they are natural for the application. In precipitation studies, the columns of **G** (the time factor) were normalized so that their mean values (or weighted means) were equal to one. Then elements of **F** represent mean concentrations or mean depositions for each compound in each factor.

Direct plotting of $\mathbf{F}$ may not be practical if different columns of $\mathbf{F}$ are of widely varying magnitudes. Several scaled-by-columns versions $\bar{\mathbf{F}}$ have been used for displaying $\mathbf{F}$ in an intuitive way:

$$\max_l \bar{\mathbf{F}}_{lj} = 1, \qquad \text{computed by setting } \bar{\mathbf{F}}_{lj} = \mathbf{F}_{lj}/\max_l \mathbf{F}_{lj}$$

$$\sum_l \bar{\mathbf{F}}_{lj} = 1, \qquad \text{computed by setting } \bar{\mathbf{F}}_{lj} = \mathbf{F}_{lj}/\sum_l \mathbf{F}_{lj}$$

$$\sum_l \bar{\mathbf{F}}_{lj} = 1 - \phi_j, \qquad \text{computed by setting } \bar{\mathbf{F}}_{lj} = (1-\phi_j)\mathbf{F}_{lj}/\sum_l \mathbf{F}_{lj}$$

where each ϕ_j is the residual variation ratio of column j, defined by

$$\phi_j = \frac{\sum_i E_{ij}^2/\sigma_{ij}^2}{\sum_i \mathbf{X}_{ij}^2/\sigma_{ij}^2}.$$

The quantity ϕ_j indicates the amount of variation of column j which is *not* explained by the factors. The quantity $1 - \phi_j$ may be called the explained weighted variation ratio of variable j. The individual scaled values $\bar{\mathbf{F}}_{lj}$ indicate how much of the variation of variable j is explained by each factor l ($l = 1, \ldots, p$). This scaling indicates the importance of each factor element $\mathbf{F}_{lj}$ in explaining the weighted variation of $\mathbf{X}$.

12. ARTIFICIAL AND REAL EXAMPLES OF PMF

The final justification for using new methods of data analysis can only come from applications. The problem with PMF is that it can be applied in quite different areas. The experience from a different area is usually irrelevant when considering the applicability of a new method in one's own research. Thus we do not report any details of applications in this paper. In the companion papers, applications of PMF in analysis of precipitation samples are described in detail, see Anttila *et al.*,[11] Tapper,[3] and Juntto and Paatero.[9]

A list of promising but as yet untried applications is given at the end of this section. The following is a list of applications of PMF which have been tried so far. Publications are in preparation for some of these applications.

1. Two-dimensional Gaussian peaks (similar to peaks occurring in coincidence spectroscopy), simulated and real data.
2. The simulation of a group of overlapping Gaussian peaks, each decaying exponentially with a unique half life. Thus on the G side, each correct unknown factor (column of G) has the shape of a Gaussian peak. On the F side, each correct row has the exponential shape. When solving the problem, the PMF model does not have any information on the special analytic forms of peaks shape or decay curve. Thus a successful analysis should find the correct shapes 'without being told'. This is in agreement with many spectroscopical situations where the exact peak shape may be unknown.
   The correct $\mathbf{X}$ is defined as the sum of four matrices of rank one,

$$\mathbf{X}_{ij}^k = \exp^{-(i-\mu_k)^2/2\omega_k^2} \exp^{-j\tau_k} \qquad \begin{matrix} k = 1, \ldots, 4 \\ i = 1, \ldots, 40 \\ j = 1, \ldots, 20. \end{matrix}$$

This example is recommended for testing factorization programs. Depending on the actual numerical values of the parameters for peak locations, widths, and decay constants $(\mu_k, \omega_k, \tau_k)$, this problem can be quite difficult numerically. On the other hand, the results are intuitively meaningful. Thus it is easy to see what is wrong with the results.
The result of this problem has rotational freedom. If no specific rotations are requested, the factors resulting from PMF are not pure peaks but combinations of them. A modified form of PMF was run so that it was made to search for such **G** factors which are as far from each other as possible. This resulted in pure Gaussian peaks and correct exponential decay curves, down to the level of simulated noise.

3. Quail Roost II simulation data (concentrations of airborne emissions), see Currie *et al.*[10] The PMF results of this example are not p-rotatable. Thus our results are uniquely determined. They are *very* close to the known correct values.
4. Data from 12 sites of the acid rain study NADP (USA) have been analysed. However, investigation of collocated sites revealed that laboratory error estimates for several compounds (especially Na, K) did not fully explain random variation of data. These elements tended to form unique factors. It will be necessary to rerun the analyses with error estimates derived from collocated sites instead of laboratory error estimates. We would like to find a research partner in USA for continuing the NADP research as a collaboration. Is anybody interested in learning how to apply the PMF on environmental data?
5. Impactor measurements of carbon-containing aerosol particles during the SCAQS campaign.
6. Analysis of origins of aerosol particles by analysing their elemental compositions. Each row of **X** consists of composition data for one particle. It is hoped that each factor would display the composition of one source and furthermore that most particles would be explained by one source only. This work is not finished yet.
7. A set of X-ray diffraction spectra, measured in different temperatures and displaying phase changes of the specimen. This problem required that rotations be controlled by hand in order to produce physically meaningful results.

In the following, some cases are mentioned which might well be tackled with PMF. We are interested in establishing co-operation with other research groups for exploring these possibilities.

(i) Determination of independently occurring components of aerosol distributions in environment. Each row of matrix **X** would contain measurements of aerosols made at one moment of time. One could pool together size distributions, chemical compositions, special properties (e.g. hydrophilic/hydrophobic, growth rate in humid conditions) etc. The more variables are available, the more resolving power one could expect from PMF. Each F factor would show the properties of one aerosol component. The corresponding G factor would indicate the concentration of that component as a function of time.

(ii) Combination of aerosol data and gaseous pollutant data. In addition to the variables mentioned above, one could pool in the matrix concentrations measured for various gaseous components. This would shed more light on the processes operating in the atmosphere. Also, it would be possible to pinpoint the sources better than with one kind of data alone.

(iii) Analysis of repeated spectroscopy or chromatography data in situations where the same compounds occur in repeated measurements so that the concentrations of different compounds vary independently. Mixtures of unknown constituents could be analysed without knowing the response of the instrument to individual constituents. It would be

possible to analyse groups of peaks, even if the individual peaks are badly overlapping; in this case ordinary peak fitting programs give unstable or unreliable results.

(iv) Analysis of data produced by two combined analytical instruments, e.g. chromatograph–optical spectrometer. The matrix **X** would consist of intensity values in a grid defined by two variables, e.g. (retention time, optical wavelength) or (retention time, molecular mass). Each pair of G and F factors would display the spectral shapes of one compound with respect to these two variables.

Many problems of this kind are being solved quite well with AR. The relatively straightforward AR algorithms have been adapted for handling extremely large matrices which typically occur in these problems, but if some cases turn out to be too difficult for AR because of difficult factor shapes, then PMF could be tried in this field, too.

13. OPEN QUESTIONS AND SUBJECTS OF FURTHER STUDY

The problem caused by multiple minima of Q needs further study. One could hope to find criteria which indicate which matrices are free of this problem or perhaps which are especially prone to this problem. A systematic approach for finding all the local minima is needed, the technique of using different sets of random numbers as starting values for the iteration need not necessarily find all the local minima of Q. It has been observed that locking a few factor components to zero often causes multiple minima of Q. A simple indicator for free rotations would be useful.

The extension of PMF approach to three matrix problems $\mathbf{X} = \mathbf{GFW} + \mathbf{E}$ where **W** is a known matrix has been reported in conference abstracts, see Paatero *et al.*[1,12] These models need proper evaluation. They will probably be useful in a variety of different problems, the matrix **W** may represent either equipment effects (i.e. Kernel) or peculiarities of the physical model, e.g. monotonicity. These models will be discussed in a forthcoming paper.

The PMF approach may also be useful in solving three-way models, e.g. in chemometrics. The extension of current algorithms to three unknown matrices should prove straightforward.

14. CONCLUSIONS

A general tool 'positive matrix factorization' or 'PMF' has been introduced for working with factor analytic or 'bilinear' models. This tool attempts to solve the quantitative problem of explaining all of the contents of the measured data matrix **X**. This problem is more ambitious than the attempt to explain correlation by FA or PCA.

It was shown that the customary PCA is obtained as a special case of PMF, omitting all non-negativity constraints and using a special set of standard deviations (all entries in one column or row having the same value, proportional to the variance of the values in the row/column) in PMF reduces PMF to the usual PCA (without the orthogonality of factors, however). PMF is able to utilize realistic error estimates of data values. This gives it several good properties: it does not generate noise factors; it handles missing values and outliers without undue loss of information, and it is expected to generate minimum variance factors because it performs an optimally weighted least squares fit. In environmental work, some degree of robustness is achieved in PMF by judicious choice of error estimates. This may not be easy for the first-time user. A period of training may be necessary in order to fully utilize the properties of PMF.

ACKNOWLEDGEMENTS

Financial support from the Maj and Tor Nessling Foundation is gratefully acknowledged. The newest version of PMF algorithm (simultaneous determination of G and F) has been

programmed by Veijo Honkimäki. The staff at the Natural Resources Ecology laboratory of Colorado State University has been most helpful in providing the NADP data.

REFERENCES

1. Paatero, P., Tapper, U., Aalto, P. and Kulmala, M. 'Matrix factorization methods for analysing diffusion battery data', *Journal of Aerosol Science*, **22**, Supplement 1, S273–S276 (1991).
2. Paatero, P. and Tapper, U. 'Analysis of different modes of factor analysis as least squares fit problems', *Chemometrics and Intelligent Laboratory Systems*, **18**, 183–194 (1993).
3. Tapper, U. 'Robust modelling of data errors in non-negative factor analysis of bulk wet deposition', Water and Environment Research Institute publications, Helsinki, Finland, 1994.
4. Henry, R. C. 'Multivariate receptor models', *in* Hopke, P. K. (ed), *Receptor Modeling for Air Quality Management*, Elsevier Science Publishers, Amsterdam, 1991.
5. Shen, J. and Israël, G. W. 'A receptor model using a specific non-negative transformation technique for ambient aerosol', *Atmospheric Environment*, **23**, 2289–2298 (1989).
6. Karjalainen, E. and Karjalainen, U., 'Mathematical chromatography – resolution of overlapping spectra in GC-MS', *Medical Informatics Europe*, **85**, (Proceedings) 572–578 (1985).
7. Karjalainen, E. and Karjalainen, U., 'Component reconstruction in the primary space of spectra and concentrations. Alternating regression and related direct methods', *Analytica Chimica Acta*, **250**, 169–179 (1991).
8. Karjalainen, E. Private communication, 1993.
9. Juntto, S. and Paatero, P. 'Analysis of daily precipitation data by positive matrix factorization', *Environmetrics*, **5**, 127–144 (1994).
10. Currie, L. A., Gerlach, R. W., Lewis, C. W., Balfour, W. D., Cooper, J. A., Dattner, S. L., De Cesar, R. T., Gordon, G. E., Heisler, S. L., Hopke, P. K., Shah, J. J., Thurston, G. D. and Williamson, H. J. 'Interlaboratory comparison of source apportionment procedures: results for simulated data sets', *Atmospheric Environment*, **18**, 1517–1537 (1984).
11. Antilla, P., Paatero, P., Tapper, U. and Järvinen, O. 'Source identification of bulk wet deposition in Finland by positive matrix factorization', *Atmospheric Environment* (1994).
12. Paatero, P. and Tapper, U. 'Use of three-matrix positive matrix factorization for analyzing environmental data', Conference 'Environmetrics', Espoo, Finland, 1992.