A GUIDE TO
POSITIVE MATRIX FACTORIZATION

Philip K. Hopke
Department of Chemistry
Clarkson University
Potsdam, NY 13699-5810

INTRODUCTION

The fundamental principle of source/receptor relationships is that mass conservation can be assumed and a mass balance analysis can be used to identify and apportion sources of airborne particulate matter in the atmosphere. This methodology has generally been referred to within the air pollution research community as *receptor modeling* [Hopke, 1985; 1991]. The approach to obtaining a data set for receptor modeling is to determine a large number of chemical constituents such as elemental concentrations in a number of samples. Alternatively, automated electron microscopy can be used to characterize the composition and shape of particles in a series of particle samples. In either case, a mass balance equation can be written to account for all m chemical species in the n samples as contributions from p independent sources

$$x_{ij} = \sum_{k=1}^{p} f_{ik} \cdot g_{kj} \qquad (1)$$

where x_{ij} is the ith elemental concentration measured in the jth sample, f_{ik} is the gravimetric concentration (ng mg^{-1}) of the ith element in material from the kth source, and g_{kj} is the airborne mass concentration (mg m^{-3}) of material from the kth source contributing to the jth sample.

There exist a set of natural physical constraints on the system that must be considered in developing any model for identifying and apportioning the sources of airborne particle mass [Henry, 1991]. The fundamental, natural physical constraints that must be obeyed are:

1) The original data must be reproduced by the model; the model must explain the observations.

2) The predicted source compositions must be non-negative; a source cannot have a negative percentage of an element.

3) The predicted source contributions to the aerosol must all be non-negative; a source cannot emit negative mass.

4) The sum of the predicted elemental mass contributions for each source must be less than or equal to total measured mass for each element; the whole is greater than or equal to the sum of its parts.

When developing and applying these models, it is necessary to keep these constraints in mind in order to be certain of obtaining physically realistic solutions.

The critical question is then what information is available to solve equation (1). It is assumed that the ambient concentrations of a series of chemical species have been measured for a set of particulate matter samples so that the x_{ij} values are always known. If the sources that contribute to those samples can be identified and their compositional patterns measured, then only the contributions of the sources to each sample need to be determined. These calculations are generally made using the effective variance least squares approach incorporated into the EPA's CMB model. However, for many locations, the sources are either unknown or the compositions of the local particulate emissions have not been measured. Thus, it is desirable to estimate the number and compositions of the sources as well as their contributions to the measured PM. The multivariate data analysis methods that are used to solve this problem are generally referred to as ***factor analysis***.

FACTOR ANALYSIS

The factor analysis problem can be visualized with the following example. Suppose a series of samples are taken in the vicinity of a highway where motor vehicles are using leaded gasoline and a steel mill making specialty steels. For these samples, measurements of Pb, Br, and Cr are made. This set of data can then be plotted in a three dimensional space as in Figure 1. A cloud of points can be observed.

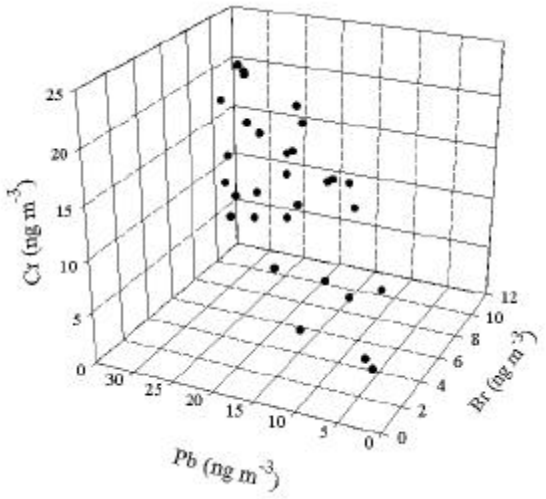

Figure 1. Three dimensional plot of simulated data.

However, it is known that there are only two particle sources. The problem is then to determine the true dimensionality of the data and the relationships among the measured variables. That is goal of a factor analysis. In the case of this example, the relationships can be observed with a simple rotation of the axes so that we look down onto the figure so that the Cr axis sticks out of the page. This view is seen in Figure 2. Now it can be seen that the data really cluster around a line that represents the Pb-Br relationship in the particles emitted by the motor vehicles. The Cr values are distributed vertically and

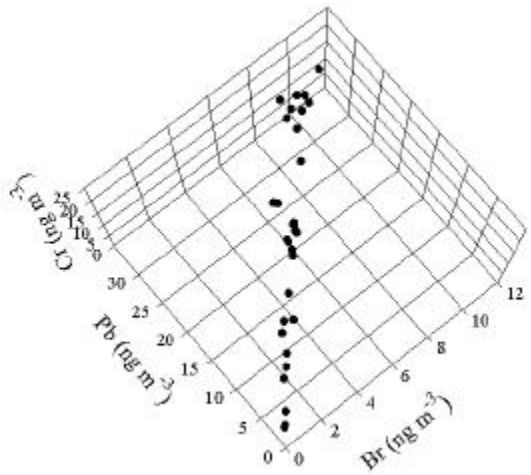

Figure 2. Plot of the simulated data as viewed from above relative to the view in Figure 1.

are independent of the other two elements. Factor analysis of this problem would find two sources and provide the relationship between the lead and bromine.

Principal Component Analysis

The most common form of factor analysis is Principal Components Analysis (PCA). This method is generally available in most computer packages for statistical analysis. The PCA results are generally calculated using an eigenvector analysis of a correlation matrix [Hopke, 1985; Henry, 1991]. However, there are several problems associated with such methods [Henry, 1987; Paatero and Tapper, 1993, 1994]. The PCA model is based on a series of n samples in which m chemical species have been measured. Thus, equation (1) can be written in matrix form as

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \tag{2}$$

The **X** matrix can also be defined in terms of the *singular value decomposition*.

$$\mathbf{X} = \mathbf{USV}' = \overline{\mathbf{U}}\,\overline{\mathbf{S}}\,\overline{\mathbf{V}'} + \mathbf{E} \tag{3}$$

where $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ are the first p columns of the **U** and **V** matrices. The **U** and **V** matrices are calculated from eigenvalue-eigenvector analyses of the **XX'** and **X'X** matrices, respectively. It can be shown [Lawson and Hanson, 1974; Malinowski, 1991] that the second term on the right side of equation (3) estimates **X** in the least-squares sense that it gives the lowest possible value for

$$\sum \sum e_{ij}^2 = \sum \sum \left(x_{ij} - \sum f_{ik} g_{kj}\right)^2 \tag{4}$$

The problem can be solved, but it does not produce a **unique** solution. It is possible to include a transformation into the equation.

$$X = G\,T\,T^{-1}\,F \tag{5}$$

where T is one of the potential infinity of transformation matrices. This transformation is called a rotation and is generally included in order to produce factors that appear to be closer to physically real source profiles.

To illustrate this problem, Figure 3 shows simulated data for ambient samples that consist of

mixtures of the soil and basalt source profiles [Currie *et al.*, 1984]. This figure shows a plot of the iron and silicon values for a series of simulated samples. There need to be two profiles to reproduce each data point, but these "profiles" could range from the original axes to any of the other pairs of lines. Because there are no zero valued source contributions, the solid lines are not the "true" profiles. The true source profiles lie somewhere between the inner solid lines that enclose all of the points and the original axes, but without additional information, these profiles cannot be fully determined. This is the problem that was presented by Henry [1987] when he described factor analysis as "ill-posed."

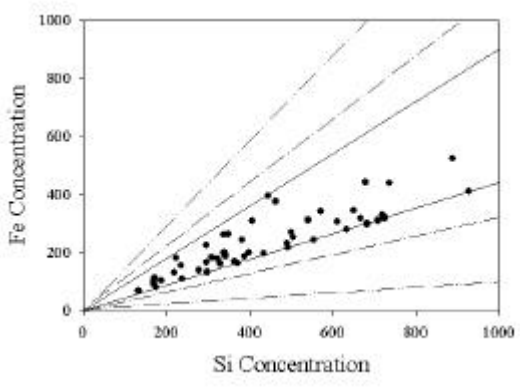

Figure 3. Simulated data showing multiple possible source "profiles" that could be used to fit the data.

Paatero and Tapper [1993] show that in effect in PCA, there is scaling of the data by column or by row and that this scaling will lead to distortions in the analysis. They further show that the optimum scaling of the data would be to scale each data point individually so as to have more precise data having more influence on the solution than points that have higher uncertainties. However, they show that point-by-point scaling results in a scaled data matrix that cannot be reproduced by a conventional factor analysis based on the singular value decomposition. Thus, an alternative formulation of the factor analysis problem is needed.

Positive Matrix Factorization

Mathematical Framework

This new approach to factor analysis is Positive Matrix Factorization (PMF). In this new method, the problem of non-optimal scaling has been explicitly addressed. In order to properly scale the data, it is necessary to look explicitly at the problem as a least-squares problem. To begin this analysis, the elements of the "Residual Matrix," E are defined as

$$e_{ij} = x_{ij} - \hat{x} = x_{ij} - \sum_{k=1}^{p} f_{ik} \cdot g_{kj} \qquad (6)$$

where i =1,..., m elements

j =1,..., n samples

k=1,..., p sources

An "object function," Q, that is to be minimized as a function of G and F is given by

$$Q(E) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{e_{ij}}{s_{ij}} \right]^2 \tag{7}$$

where s_{ij} is an estimate of the "uncertainty" in the ith variable measured in the jth sample. The factor analysis problem is then to minimize Q(E) with respect to G and F with the constraint that each of the elements of G and F is to be non-negative.

Initially the problem was solved iteratively using alternating least squares [Paatero and Tapper, 1993]. In alternating least-squares, one of the matrices, G or F, is taken as known and the chi-squared is minimized with respect to the other matrix as a weighted linear-least-squares problem. Then the roles of G and F are reversed so that the matrix that has just been calculated is fixed and the other is calculated by minimizing Q. The process then continues until convergence. However, this process can be slow.

In order to improve performance, each step in the iteration was modified so that modifications are made to both matrices [Paatero and Tapper, 1994]. Subsequently, Paatero [1997] describes a global optimization scheme in which the joint solution is directly determined. This program, PMF2, is able to simultaneous vary the elements of G and F in each iterative step. The model will be designated as PMF while the program used to solve the model will be designated PMF2.

In PMF2, the non-negativity constraint is imposed through the use of a penalty function in an enhanced function, $\overline{Q}$. Regularization can also be imposed to reduce rotational freedom so that the enhanced object function is defined as

$$\begin{aligned} \overline{Q}(E,G,F) &= Q(E) + P(G) + P(G) \\ &\quad + R(G) + R(F) \\ &= \sum_{i=1}^{m} \sum_{k=1}^{n} \left[\frac{e_{ij}}{\sigma_{ij}} \right]^2 \\ &\quad - \alpha \sum_{i=1}^{m} \sum_{k=1}^{n} \log g_{ih} - \beta \sum_{i=1}^{m} \sum_{k=1}^{n} \log f_{hj} \\ &\quad + \gamma \sum_{i=1}^{m} \sum_{k=1}^{n} g_{ih}^2 + \delta \sum_{i=1}^{m} \sum_{k=1}^{n} f_{ih}^2 \end{aligned} \tag{8}$$

P(G) and P(F) are penalty functions that prevent the factors G and F from becoming negative. R(G) and R(F) are regularization functions that remove some of the rotational indeterminacy. For the unconstrained case, " and ß are set to zero. However, the regularization terms are always present in the analysis. All of the coefficients, ", $, (, and *, are given smaller values during the iterations so that their final values are negligible but not zero. In practice, the log functions are approximated by a Taylor series expansion up to quadratic terms. The algorithm can obtain transient negative values, but it immediately adjusts the Taylor expansion so that the user never sees the negative values. Only non-negative values are stored in factor matrices.

Subsequently, it was found useful to have a more versatile program for solving a more general set of problems. Xie *et al.* [1999b] found that a model that includes both terms like equation (1) and trilinear terms (product of 3 factors) provided a better fit to the data. Paatero [1999] has developed a second program, the Multilinear Engine (ME), that can also solve the positive matrix factorization problem but with an approach that provides more flexibility in specifying the model. There are differences in the computational approach between PMF2 and ME. At this time, PMF2 has been more widely used and tested and will be the focus of this discussion.

Estimation of Weights

The solution to the PMF problem depends on estimating uncertainties for each of the data values used in the PMF analysis. There are three types of values that are typically available. Most of the data points have values that have been determined, x_{ij}, and their associated uncertainties, σ_{ij}. There are samples in which the particular species cannot be observed because the concentration is below the method detection limit. Finally there are samples for which the values were not determined. These latter two types of data are often termed "missing" data. However, there are qualitative differences between them. In the below detection limit samples, the value is known to be small, but the exact concentration is not known. In the case where values could not be determined, the value is totally unknown.

Polissar *et al.* [1998] has suggested the following for the concentration values and their associated error estimates:

$$\begin{aligned} x_{ij}^k &= v_{ij}^k && \text{For determined values} \\ x_{ij}^k &= d_{ij}^k/2 && \text{For below limit of detection values} \\ x_{ij}^k &= \tilde{v}_{ij}^k && \text{For missing values} \end{aligned} \qquad (9)$$

$$\sigma_{ij}^{k} = u_{ij}^{k} + d_{ij}^{k}/3 \quad \text{For determined values}$$
$$\sigma_{ij}^{k} = \bar{d}_{ij}^{k}/2 + d_{ij}^{k}/3 \quad \text{For below limits of detection values} \qquad (10)$$
$$\sigma_{ij}^{k} = 4\tilde{v}_{ij}^{k} \quad \text{For missing values}$$

where v_{ij}^{k}, u_{ij}^{k}, and d_{ij}^{k} are the measured concentration, the analytical uncertainty, and the method detection limit, respectively, for sample *i*, element *j*, and sampling site k, $\bar{d}_{j}^{k}$ is arithmetic mean of the detection limit value for the element *j* and the sampling site *k*; and $\tilde{v}_{j}^{k}$ is the geometric mean of the measured concentration of the element *j* at sampling site *k*. Half of the average detection limits were used for below detection limits values in the calculations of the geometric means.

The detection limits d_{ij}^{k} specified the error estimates for low concentration v_{ij}^{k} while uncertainties u_{ij}^{k} provided the estimation of errors for high concentration v_{ij}^{k} of element *j*. Error estimates for the determined values were usually less than 50%; $d_{ij}^{k}/2$ were used for below detection limits values and corresponding relative error estimates from 100% to 250% were used for these values according to equation (10). Geometric mean values with the corresponding error estimates equal to 4 times these values were used for the missing values so that relative error estimates for missing values were equal to 400%. Thus, both the $d_{ij}^{k}/2$ values that were used for the below detection limit values and the $\tilde{v}_{ij}^{k}$ values that were used for missing values had much lower weights in comparison to actually measured values.

"Missing" means that some elemental concentrations are missing for a specific sample but the concentration of at least one element had been reported for the particular sample. Then the missing concentrations and corresponding error estimates of other elements were calculated according to equations (9) and (10). Alternatively, PMF2 can compute heuristic error estimates, s_{ij}, for x_{ij} based on the data point and its original error estimate, F_{ij}. The general equation is

$$s_{ij} = C_1 + C_2 \sqrt{|x_{ij}|} + C_3 * x_{ij} * \qquad (11)$$

with

C_1 = F_{ij}

C_2 = 0 except for Poissonian processes

C_3 = dimensionless values between 0.1 – 0.2 were chosen so that the relative uncertainty of each data set was reasonable.

Similar approaches were used for estimating weights in other recent applications [Yakovleva *et al.*, 1999; Chueinta *et al.*, 2000].

PMF also offers an option when it is not possible to fully specify the individual data point errors. Models can be developed for this problem using a distribution that is known *a priori*. For environmental data, a lognormal model is available in which the user has to specify the logarithm of the "geometrical standard deviation" (log(GSD)) for each measured value. It is often sufficient to use the same value for all measured points if all of the values are truly lognormally distributed. However, zero values cannot occur in a lognormal distribution and yet such values can occur. These values make the factorization unreliable if they are processed under the assumption of pure lognormality. Then the assumption of lognormality needs to be modified. PMF offers the option to assume that there is an additional normal error superimposed onto the lognormal distribution. The method detection limit can be used as this additional error. The statistical properties of this model are not fully known. The expectation value of Q does not equal the degrees of freedom in lognormal model. However, PMF provides an estimate of the expected value of Q

Another important aspect of weighting of data points is the handling of extreme values. Environmental data typically shows a positively skewed distribution and often with a heavy tail. Thus, there can be extreme values in the distribution as well as true “outliers.” In either case, such high values would have significant influence on the solution (commonly referred to as leverage). This influence will generally distort the solution and thus, an approach to reduce their influence can be a useful tool. Thus, PMF offers a “robust” mode. The robust factorization based on the Huber influence function [Huber, 1981] is a technique of iterative reweighing of the individual data values. The least squares formulation, thus, becomes to

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} (e_{ij} / h_{ij} s_{ij})^2 \qquad (12)$$

where

$$h_{ij}^2 = \begin{cases} 1 & \text{if } \left| e_{ij} / s_{ij} \right| \le \alpha, \text{ and} \\ \left| e_{ij} / s_{ij} \right| / \alpha & \text{otherwise} \end{cases}$$

" = the outlier distance and the value of " = 4.0 was chosen

It is generally advisable to use the robust mode when analyzing environmental data.

Our experience has generally found that the lognormal model and the robust mode provide the best results for typical particulate composition data. For the NOAA aerosol data from Barrow [Polissar *et al.*, 1999], it was found that the lognormal model produced reasonable results but that the robust mode caused problems in terms of results that made physical sense. However, these data consist of monthly mean particle counts and light scattering efficiency and apparently behave somewhat differently from typical particle composition data.

Algorithms

Over the past several years several approaches to solving the PMF problem have been developed. Initially, a program called PMF2 utilized a unique algorithms [Paatero, 1997] for solving the factor analytic task. For small and medium-sized problems, this algorithm was found to be more efficient than ALS methods [Hopke *et al.*, 1998]. Subsequently, an alternative approach that provides a flexible modeling system has been developed for solving the various PMF factor analytic least squares problems [Paatero, 1999]. This approach, called the multilinear engine (ME), has been applied to an environmental problem [Xie *et al.*, 1999a], but has not yet been widely used.

Estimation of the Number of Factors

The theoretical Q-value should be approximately equal to the number of degree of freedom, or approximately equal to the total number of data points in the data array. If the errors are properly estimated, then it can be seen that fitting each data point such that the reproduced value to within the estimated error value will contribute a value of approximately 1 to the Q value. Thus, it is tempting to examine the estimated Q value as a function of the number of factors to determine the number of factors to retain. However, this approach can be misleading if the data point uncertainties are not well determined. This problem has been observed when only the measured uncertainties are used as the weights. This problem can be minimized by using weights like are presented in Equation 10 or using a model for the data point errors as described in Equation 11. It is useful to look at the changes in Q as additional factors are calculated since after an appropriate number of factors are included in the fit, additional factors will not result in significant further improvements in the Q-value. Such behavior can be observed in Yakovleva *et al.* [1999]. Thus, PMF suffers from the same difficulties in determining the correct number of factors as all other forms of factor analysis.

As one of the methods to determine the quality of the fit, the residuals, e_{ij}, are examined. Typically the distributions of the residuals are plotted for each measured species. It is desirable to have symmetric distributions and to have all of the residuals within $\pm$ 3 standard deviations. If there is a larger spread in the residuals, then the number of factors should be reexamined.

One of the disadvantages of a least-squares approach is that it can yield multiple solutions depending on the initial starting point. As part of the PMF program, it is possible to initiate random values in the F and G matrices as the starting point for an analysis. It is generally advisable to perform the analysis several times (typically 5) to be certain that the same solution is obtained. It has been our experience that one of the tests of the best selection of the number of factors is that one does not obtain multiple solutions or obtains at most one alternative solution. With greater or fewer factors than optimum, multiple solutions are more often obtained.

Factor Rotations

In general, this type of bilinear factor analysis will have rotational ambiguity. There will not be a unique solution even though there will be a global minimum in the least-squares fitting process. The addition of constraints can reduce the rotational freedom in the system, but non-negativity alone does not generally result in a unique solution. One of the key features of PMF is that the rotations are part of the fitting process and not applied after the extraction of the factors as is done in eigenvector-based methods. Rotational freedom can be reduced by having known zero values in either the **G** or **F** matrices and forcing the solution to those values. Since rotations really represent addition and subtractions [Paatero and Tapper, 1994], the combination of zero values and non-negativity constraints reduces the rotational ambiguity.

One of the ways that rotations can be controlled in PMF2 is by using the parameter FPEAK. By setting a non-zero value of FPEAK, the routine is forced to add one G vector to another and subtract the corresponding F factors from each other and thereby yield more physically realistic solutions. The method requires a complicated mathematical process which is not described here [see Paatero, 1997]. The degree of rotational freedom is reduced if there are real zeroed values in the true F and G matrices. Then for positive FPEAK, zeroes in the F matrix restrict subtractions since no resulting value can become negative. Alternatively, a negative FPEAK causes subtractions in the G matrix and those subtractions are restricted if there zeroes in the matrix. If there are a sufficient number of zero values both in **F** and in **G**, then the solution is rotationally unique: there are no possible rotations. However, this case is rare.

Experience for real data sets has shown that positive, non-zero FPEAK values generally yield more realistic results. There is no theoretical basis for choosing a particular value of FPEAK. However, it is important to examine the Q value. Since the rotation is integrated into the optimization scheme, the Q can change with the rotation. If the rotation forces the fit too away from the original Q, then the value of FPEAK should be moved toward zero. Figure 5 shows the behavior that has been observed for a number of data sets when Q is examined as a function of FPEAK. Typically the highest FPEAK value before the substantial rise in Q yielded the most physically interpretable source profiles. It has also been found in several cases that when examining the regression of the mass against the G values, FPEAK values can be varied to yield non-negative scaling factors while retaining physically reasonable source profiles.

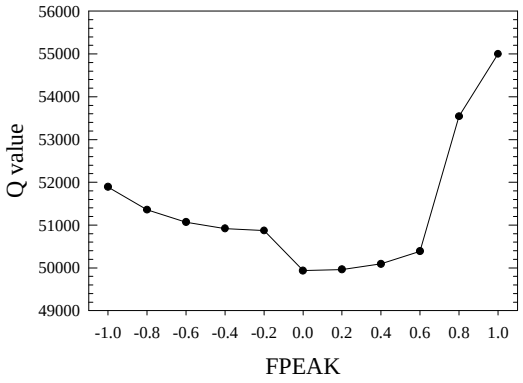

Figure 5. Plot of Q as a function of FPEAK for a fine particle data set from Phoenix, AZ.

The imposition of external information on the solution is another way in which PMF can control rotations. If specific values of profiles or time series are known to be zero, then it is possible to force the solution toward zero for those values through appropriate settings of Fkey and Gkey values. Details of setting up the input files are in the PMF User Guide.

When analyzing Hong Kong aerosol, Lee et al. found sulphate in almost all factors (Lee *et al.*, 1999). Such compositions of factors did not appear plausible. The concentration of sulphate was then pulled down in a number of factors in another PMF2 run. The increase of Q was not significant and more plausible factors resulted from the pulled-down computations.

The analysis of Hong Kong data offers a simple example of pulling-down: only one compound is subjected to pulling. It would be possible to use pulling on several compounds, provided that there is reliable information about compounds that are not emitted by certain sources. In addition, it might also be possible to apply pulling-down to time series factors. Sometimes information about weather patters might reveal that a certain source cannot be present during certain time intervals. Then the time factor elements corresponding to such intervals could be pulled towards zero.

Error Estimates for the PMF Results

It is useful to have an estimate of the uncertainties in the elements of the **F** and **G** that have been calculated using PMF2. It is possible to estimate uncertainties using the process originally described by Roscoe and Hopke [1981] and described in detail by Malinowski [1991]. The errors in the elements of one matrix are estimated based on the errors in the ambient concentration values and assuming that the other matrix is error-free. Each matrix (**F** or **G**) is treated similarly in turn such that an uncertainty is associated with each matrix element. Error models are currently under development that may produce better estimates for the PMF analysis results.

Mass Apportionment

The results of the PMF analysis reproduce the data and ensure that the source profiles and mass contributions are non-negative. However, it has not yet taken into account the measured mass. In addition the results are uncertain relative to a multiplicative scaling factor. Again a "1" can be introduced into equation 1.

$$x_{ij} = \sum_{k=1}^{p} f_{ik} \cdot g_{kj} = \sum_{k=1}^{p} f_{ik} \cdot \frac{s_k}{s_k} \cdot g_{kj} \qquad (10)$$

If we have a sufficiently detailed chemical analysis such that those species that were not measured can either be assumed to be strongly correlated to measured species or represent sources that add negligible mass to the particulate matter samples, then the sum of the source contributions, g_{kj} values, should be equal to the measured PM mass. Using this external information, the scaling constants, s_k, can be determined by regressing the measured mass against the estimated source contribution values.

$$m_j = \sum_{k=1}^{p} s_k \cdot g_{kj} \qquad (11)$$

This regression provides several useful indicators of the quality of the solution. Clearly each of the s_k values must be non-negative. If the regression produces a negative value, then it suggests that too many factors have been used. The source profiles need to be scaled by dividing the f_{ik} value by s_k. Once the profiles are scaled, they can be summed and it can be determined if the sum of a source profile exceeds 100% (1000 ng/∶g). In this case, it suggests that too few factors may have been chosen. Although it is desirable to require that the s_k values be statistically significant, there can be weak or aperiodic sources for which these scaling coefficients may have high uncertainties.

APPLICATIONS

PMF2 was initially applied to data sets of major ion compositions of daily precipitation samples collected over a number of sites in Finland [Juntto and Paatero, 1994] and samples of bulk precipitation [Anttila *et al.*, 1995] in which they are able to obtain considerable information on the sources of these ions. Polissar *et al.* [1996] applied the PMF2 program to Arctic data from 7 National Park Service sites in Alaska as a method to more quantitatively resolve the major source contributions.

Recently there has been a series of applications of PMF to various source/receptor modeling problems. Polissar *et al.* [1998] have reanalyzed an augmented set of Alaskan NPS data and resolved up to 8 sources. Xie *et al.* [1999a&b] have made several analyses of data from an 11 year series of particulate matter samples taken at Alert, N.W.T. Polissar *et al.* [1999] have examined the semicontinuous aerosol data collected by NOAA at their atmospheric observatory at Barrow, Alaska.. Lee *et al.* [1999] have applied PMF to urban aerosol compositions in Hong Kong. They were able to identify up to 9 sources that provided a good apportionment of the airborne particulate matter. Paterson *et al.* [1999] applied PMF to air quality and temperature data collected at a series of sites around the southern end of Lake Michigan in 1997 and used three factors to reproduce 75% of the variation in the data. Huang *et al.* [1999] analyzed elemental composition data for particulate matter samples collected at Narragansett, R.I. using both PMF and conventional factor analysis. They were able to resolve more components with more physically realistic compositions with PMF. Thus, the approach is gaining interest because it does have some inherent advantages particularly through its ability to individually weight each data point. PMF is somewhat more complex and harder to use, but it appears to provide improved resolution of sources and better quantification of impacts of those sources than PCA [Huang *et al.*, 1999].

ADVANTAGES OF PMF

There are substantial advantages to individual data point weights. Data point weights permit the adjustment of the influence of the values to be related to the level of confidence the analyst has in the data. In the case of the NPS data from Denali, Alaska, Polissar *et al.* [1998] were able to extract a good sea salt profile from the data even though more than 80% of the Cl values are below the detection limits. In general, more of the data can be retained in the analysis than is possible with eigenvector-based approaches where such a high level of “missing” data would have resulted in badly distorted results.

The least-squares approach also permits the inclusion of additional constraints as an integral part of the analysis. Wang and Hopke [1989] originally showed the value of constrained least-squares fits in

receptor models. The PMF2 program that implements PMF includes non-negativity constraints, but it is not written in such a manner that additional constraints can be applied to the problem. Additional constraints can be imposed using the multilinear engine (ME) program and it can also provide the other options available for PMF. However, it is currently more complicated to use as one has to script the constraint model for the problem. It would be possible to produce a preprocessing program that would prepare the input files for the ME program that could incorporate all of the constraints that are appropriate for any given problem.

Thus, the PMF model can provide inherently better modeling of the data than eigenvector-based methods. It allows individual data point weights that permits maximum use of the available data and as implemented in the ME program, permits the imposition of all of the constraints that can be applied using any other approach. Because of the number of options and the requirement to develop the input weight matrix, it is more complex to use either PMF2 or ME programs. As a research tool, this complexity has not been a problem, but to be used on a more routine basis, the programs will need to be made easier to use.

COMPARISON WITH CMB

Both CMB and PMF provide quantitative estimates of the source contributions. In the CMB analysis, source profiles are provided whereas in PMF, the source profiles are estimated. If it is some of the source profiles are known, they can be used in PMF to constrain the extracted source profiles and thereby reduce the rotational indeterminacy. Both CMB and PMF are employing least squares fitting, but there are some important differences in how the underlying error structures are modeled and how many unknowns are being estimated. With PMF it is not possible to precisely assign errors to the source profiles and contributions. In a CMB analysis, it is possible to assign define error estimates to each source contribution value. However, there are no diagnostics provided in the CMB model that would alert the user to misspecification of the source profiles. Also since the CMB analysis is done on a sample-by-sample basis, there can be errors in the estimated source contributions because of the variations that can occur in the source profiles. PMF uses all of the data and thus, estimates the average source profile over the time interval during which samples were acquired. Thus, there are some similarities in the process and the outcome, but there are also some important differences in what is being estimated, the input data that is required, and the estimates of the uncertainties in the calculated values.

REFERENCES

Anttila, P., P. Paatero, U. Tapper, and O. Järvinen (1995) Application of Positive Matrix Factorization to Source Apportionment: Results of a Study of Bulk Deposition Chemistry in Finland, *Atmospheric Environ.* 29:1705-1718.

Chueinta, W., P.K. Hopke and P. Paatero (2000) Investigation of Sources of Atmospheric Aerosol Urban and Suburban Residential Areas in Thailand by Positive Matrix Factorization, *Atmospheric Environ.* In press.

Currie, L.A., R.W. Gerlach, C.W. Lewis, W.D. Balfour, J.A. Cooper, S.L. Dattner, R.T. DeCesar, G.E. Gordon, S.L. Heisler, P.K. Hopke, J.J. Shah, G.D. Thurston, and H.J. Williamson, Interlaboratory Comparison of Source Apportionment Procedures: Results for Simulated Data Sets, *Atmospheric Environ.* 18:1517-1537 (1984).

Henry, R.C. (1987) Current Factor Analysis Models are Ill-Posed, *Atmospheric Environ.* 21:1815-1820.

Henry, R.C. (1991) Multivariate Receptor Models, In: *Receptor Modeling for Air Quality Management*, P.K. Hopke, ed., Elsevier Science Publishers, Amsterdam, 117-147.

Hopke, P.K. (1985) *Receptor Modeling in Environmental Chemistry*, John Wiley & Sons, Inc., New York.

Hopke, P.K., ed. (1991) *Receptor Modeling for Air Quality Management*, Elsevier Science, Amsterdam.

Hopke, P.K.,P. Paatero, H. Jia, R.T. Ross, and R.A. Harshman (1998) Three-Way (PARAFAC) Factor Analysis: Examination and Comparison of Alternative Computational Methods as Applied to Ill-Conditioned Data, *Chemom. Intel. Lab. Syst.* 43:25-42.

Huang, S., K.A. Rahn and R. Arimoto (1999) Testing and Optimizing Two Facot-Analysis Techniques on Aerosol at Narragansett, Rhode Island, *Atmospheric Environ.* 33:2169-2185.

Huber, P. J. (1981) *Robust Statistics*, John Wiley, New York.

Juntto, S. and P. Paatero (1994) Analysis of daily precipitation data by positive matrix factorization, *Environmetrics* 5:127-144.

Lawson, C.L. and R.J. Hanson (1974) *Solving Least-Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ.

Lee, E., C.K. Chan, and P. Paatero (1999) Application of Positive Matrix Factorization in Source Apportionment of Particulate Pollutants in Hong Kong, *Atmospheric Environ.* 33:3201-3212.

Malinowski, E.R. (1991) *Factor Analysis in Chemistry*, Wiley, New York, 2nd Ed..

Paatero, P. (1997) Least Squares Formulation of Robust, Non-Negative Factor Analysis, *Chemom. Intell. Lab. Syst.* 37:23-35.

Paatero, P. (1999)The Multilinear Engine --- a Table-driven Least Squares Program for Solving Multilinear Problems, Including the n-way Parallel Factor Analysis Model, J. Computational and Graphical Stat. 8: 1-35.

Paatero, P. and U. Tapper (1993) Analysis of Different Modes of Factor Analysis as Least Squares Fit Problems, *Chemom. Intell. Lab. Syst.* 18:183-194.

Paatero, P. and U. Tapper (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* 5:111-126.

Paterson, K.G. , J.L. Sagady, D.L. Hooper , S.B. Bertman , M.A. Carroll , and P.B. Shepson (1999) Analysis of Air Quality Data Using Positive Matrix Factorization, *Environ. Sci. Technol.* 33: 635-641

Polissar, A.V., P.K. Hopke, W.C. Malm, J.F. Sisler (1996) The Ratio of Aerosol Optical Absorption Coefficients to Sulfur Concentrations, as an Indicator of Smoke from Forest Fires when Sampling in Polar Regions, *Atmospheric Environ.* 30:1147-1157.

Polissar, A.V., P.K. Hopke, W.C. Malm, J.F. Sisler (1998) Atmospheric Aerosol over Alaska: 2. Elemental Composition and Sources, *J. Geophys. Res.* 103: 19,045-19,057.

Polissar, A.V., P.K. Hopke, P. Paatero, Y. J. Kaufman, D. K. Hall, B. A. Bodhaine, E. G. Dutton and J. M. Harris (1999) The Aerosol at Barrow, Alaska: Long-Term Trends and Source Locations, *Atmospheric Environ.* 33: 2441-2458.

Roscoe, B.A. and P.K. Hopke (1981) Error Estimation of Factor Loadings and Scores Obtained with Target Transformation Factor Analysis, *Anal. Chim. Acta* 132:89-97.

Wang, D. and P. Hopke (1989) The Use of Constrained Least-Squares to Solve the Chemical Mass Balance Problem, *Atmospheric Environ.* 23:2143-2150.

Xie, Y.-L., P. K. Hopke, P. Paatero, L. A. Barrie and S.-M. Li (1999a) Identification of Source Nature and Seasonal Variations of Arctic Aerosol by the Multilinear Engine, *Atmospheric Environ.* 33: 2549-2562.

Xie, Y. L., Hopke, P., Paatero, P., Barrie, L. A., and Li, S. M. (1999b). Identification of source nature and seasonal variations of Arctic aerosol by positive matrix factorization. *J. Atmos. Sci.* 56:249-260.

Yakovleva, E., P.K. Hopke and L. Wallace (1999) Receptor Modeling Assessment of PTEAM Data, *Environ. Sci. Technol.* 33: 3645-3652 (1999).