JRC Scientific and Technical Reports

Positive Matrix Factorisation (PMF)

An introduction to the chemometric evaluation of environmental monitoring data using PMF

Sara Comero, Luisa Capitani and Bernd Manfred Gawlik

EUR 23946 EN - 2009

EUROPEAN COMMISSION

Institute for Environment and Sustainability

The mission of the JRC-IES is to provide scientific-technical support to the European Union's policies for the protection and sustainable development of the European and global environment.

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information
Address: Via E. Fermi, 2479 I-21020 Ispra (VA) Italy
E-mail: sara.comero@jrc.europa.eu
Tel.: +39 0332 78 5935
Fax: +39 0332 78 9831

http://ies.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

Legal Notice
Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

JRC 52754

EUR 23946 EN
ISBN 978-92-79-12954-4
ISSN 1018-5593
DOI 10.2788/2497

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Positive Matrix Factorisation (PMF)

An introduction to the chemometric evaluation of environmental monitoring data using PMF

Sara Comero, and Bernd Manfred Gawlik
European Commission Joint Research Centre, Institute for Environment and Sustainability,
21020, Ispra, Italy

Luisa Capitani
Department of Earth Sciences "*Ardito Desio*", Section of Geochemistry and Volcanology.
Università degli Studi di Milano, via Mangiagalli 34, 20133 Milano Italy

Abstract

Positive Matrix Factorization (PMF) is a multivariate factor analysis technique used successfully among others at the US Environmental Protection Agency for the chemometric evaluation and modelling of environmental data sets. Compared to other methods it offers some advantage that consent to better resolve the problem under analysis. In this report, the algorithm to solve PMF and the respective computer application, PMF2, is illustrated and, in particular, different parameters involved in the computation are examined.

Finally, a first application study on PMF2 parameters setting is conducted with the help of a real environmental data-set produced in the laboratories of the JRC Rural, Water and Ecosystem Resource Unit.

Table of contents

List of Figures

1 Introduction and rationale

Receptor models are used in different branches of scientific research (i.e. atmospheric and geochemical research) because of they capability to handle large data-sets. The aim of their application is to reduce the original data-set into one of lower dimensions to detect “hidden” information and explain the variability of the measured variables. In particular, in environmental applications the goal of receptor modelling is to estimate number and composition of sources (the factor that explains the data variability) but also to point out any trend and/or correlation among observations and identify potential marker for pollutant sources.

Among the different type of available receptor models (see Chapter 2), in this paper we want to focus the attention on the model PMF (Positive Matrix Factorization).

The reason we are interested in PMF is its property to be a non data-sensitive technique that can manage and resolve inhomogeneous data-sets without any previous univariate analysis. In geochemical data-set, it usually happens that elements or compounds occurring in very different concentrations caused, for examples, by the presence of different geochemical features. This may be a problem for data-sensitive techniques and normalization procedures have to be applied to homogenize the original data-set. This, however, causes loss of information.

Another important aspect of PMF is the introduction of error estimates (or weight) associated to the data. Like this, problematic data such as outliers or below-detection-limit can be entered into the model with appropriated weight, avoiding rejection of such data.

In Chapter 3 we give an introduction to PMF model explaining the mathematical algorithm and the selection of appropriate error estimates, while in the Appendix a guide to the program PMF2 used to solve PMF (Paatero, 2004a; Paatero 2004b) is given.

Finally, in the last chapter, we describe the main steps to interpret results produced by PMF2 and the use of the parameters involved in the computation to obtain a solution that better describe the real problem; this will be done with the help of a practical example. This report is the basis for further PMF research in the framework of the PhD “*Fate assessment and source apportionment of environmental pollutants using X-ray* *analytical techniques and chemometric data modelling*", which objective is to develop a PMF toolbox for different kind of environmental monitoring data-sets. The toolbox will be use to extract "hidden" data structures (i.e. regional geomorphology) and to identify markers for pollution compounds or sources by a chemometric approach.

2 Chemometric modelling

2.1 *Development in chemometric modelling*

Environmental monitoring data are more and more often handled in terms of mathematical models that allow managing different kind of dataset with multiple observations. Depending on the type of known information (input data) and on the type of results that one would obtain (output data), different modelling technique are available.

In the recent years receptor models became an increasingly important instrument in environmental sciences in order to elicit information on dataset containing a number of features (chemical or physical properties) related to the measured samples. In particular they are used to evaluate the contamination and pollutant sources contributions in different kind of samples, starting from the information carried out by the samples (registered at monitoring site) and hence at the point of impact, or receptor. Because of this property they are diagnostic models (Hopke, 2003). Receptor models complement the source-oriented dispersion models, which are prognostic model. They are indeed based on sources emission inventory to infer sources emission effects on pollutants concentration and required a high level of information about the diffusion parameter of the problem under analysis.

Receptor models are also known as multivariate methods, because they are used to analyze data set involving a number of numerical values as a whole; they had been primarily implemented on atmospheric datasets (characterization of air pollution; Lee *et al.*, 1999; Xie *et al.*, 1998), but more and more these techniques are also used to study samples from aquatic and terrestrial compartments (DelValls *et al.*, 1998; Reimann *et al.*, 2002).

Receptor models are based on the **Chemical Mass Balance** (CMB) equations that, considered a single sample taken at a single location and time period, can be expressed as (Watson *et al.*, 2008):

$$C_{ij} = \sum_{i} F_{ik} S_{kj} + E_{ij} \qquad (2.1.1)$$

where C_{ij} is the amount of the i^{th} variable (i.e. a chemical element or compound concentration, or a physical property amount) measured at the location (sample) j; F_{ik} is the fractional abundance of the i^{th} variable in the k^{th} source type and S_{kj} is the contribution of the k^{th} source at the location j. E_{ij} represents the residuals, that is the difference between the measured and calculated amounts.

In order to obtain physically realistic solutions of the last equation, Hopke, 2000, identified some natural constraints that the system must satisfy:

1. The original data must be reproduced by the model and the model must explain the observations.
2. The predicted factor explaining the source composition must be non-negative since a negative amount does not have a physical sense (a source cannot be composed by a negative variable amount, otherwise it is a sink).
3. The predicted factor explaining the source contributions must all be non-negative since a source emitting a negative amount is physically not realistic.
4. Only for chemical elements or compounds, where the unit of measurement are the same, 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 (only in the case of chemical elements or compounds).

2.2 *Recent type of receptor models*

Starting from Equation 2.1.1 there are different ways to find a solution, depending on the type of available information and on the desired final result. Receptor models are categorized in two principal classes, based on the same Equation 2.1.1: chemical mass balance models and multivariate models.

In case of chemical mass balance models, the number and characterization of the main sources must be known *a priori*; these kinds of models are generally used in environmental studies related to the determination of the sources mass contribution, starting from the sources characterization (sources profile). Sources profiles are customary known from preceding studies or extracted from existing data set. The most useful model in this class is the Chemical Mass Balance (CMB) and the equation is solved using weighted least square regression analysis. In the majority of the cases, however, main sources are not well-known and/or inappropriate source profiles from other location are used, making the model usually very inaccurate.

To elicit information on sources type, number and contribution starting from observations (i.e. element concentrations data set) at receptor site, different factor analysis methods (multivariate methods) have been developed. Commonly factor analysis methods used in physical and chemical sciences are: Principal Component Analysis (PCA), Unmix, Target Transformation Factor Analysis (TTFA), Positive Matrix Factorization (PMF) and Multilinear Engine (ME).

Actually, PCA is referred to several forms of eigenvector analysis that have the same basic objective: the compression of data into fewer dimensions (or factors) and the identification of the correlation between the measured variables.

It should be noted that the term "*Factor Analysis (FA)*" has an ambiguous meaning to identify the above factor models (Paatero and Tapper, 1994). In fact, in statistics, it means a non-linear analysis based on investigation of correlations of random variables, which is seldom used in physics or chemistry. This factor analysis is also named "orthodox FA" or "non-linear FA" in order to distinguishing from the above listed factor analysis methods.

2.3 *PCA and Single Value Decomposition*

Since Principal Component Analysis (PCA) is one of the most commonly used method used for data analyses in environmental sciences, particularly in atmospheric research and climate, it is necessary to shortly describe this model. PCA has been used in different kind of studies related to the atmosphere, like: air pollution (Motelay-Massei *et al.*, 2003; Pires *et al.*, 2008), ozone (Yu *et al.*, 2000; Chang *et al.*, 2009) and precipitation (Sakihama *et al.*, 2008). Examples of other applications in aquatic and terrestrial compartments are known, too for instance river and lake sediments (dos Santos *et al.*, 2004; Loska *et* Wiechuła, 2003) and sewage water (Critto *et al.*, 2003) and in land studies (Officer *et al.*, 2004) and so forth..

This technique and its variants attempt to reduce the initial set of variables into a new set of casual factors with reduced dimension, by means of correlations between the measured variables.

Traditionally the factorization of PCA has been based on the so-called covariance matrix; later also the Singular Value Decomposition (SVD) has been used (Paatero, a). The first resolving algorithm requires that the data matrix be first centred, but this results in a loss of information about the origin of the scale of variables; thus this approach is inappropriate for instance in physical sciences.

Singular Value Decomposition is a matrix factorization technique that factors a given matrix X, of dimension *mxn*, into three matrices as follows:

$$X_{mxn} = U_{mxr} S_{rxr} V_{rxn}^{T} \qquad (2.3.1)$$

where U and V are orthonormal matrices ($U^T U = V^T V = I$). S is a diagonal matrix containing the *singular values* of the matrix X. There are exactly *r* singular values that correspond to the square roots of XX^T or X^TX eigenvalues, where *r* is the rank of X. With this decomposition we can also identify XX^T and X^TX eigenvectors that correspond respectively to U and V columns (Unonius and Paatero, 1990).

The SVD algorithm hence consists of finding the eigenvalues and eigenvectors of XX^T and X^TX.

Since the covariance matrix of the X is a multiple of XX^T, the SVD is able to find its eigenvectors, also called *principal component*; moreover its eigenvalues (from S matrix) are the variance associate to each principal component. Hence, selecting only the more important components (those with higher eigenvalues), say the first *h*, data are projected from *m* to *h* dimensions.

One important result of the SVD of X is that the truncated SVD of X up to an *h* order:

$$X^{(h)} = \sum_{k=1}^{h} u_k s_k v_k^T \qquad (2.3.2)$$

minimizes the sum of squares of the difference between X and $X^{(h)}$ elements. In other words, the truncated SVD forms an un-weighted least squares fit of X, giving a best possible approximation of X when the approximating matrices are restricted to *h* columns (Paatero and Tapper, 1994).

In the case of PCA the factorization is given by X≈GF where the *nxm* X matrix is the measured data set (i.e. elements concentration in different sampling sites) and G and F, respectively with *nxp* and *pxm* dimensions, are the two factor matrices explaining the resolved factors.

The X matrix can also be viewed as a sum of the matrix components of rank one $X^1+\ldots+X^p$ and applying to them the SVD, the solution of rank *p* is given by $X=\tilde{U}\tilde{S}\tilde{V}'+E=GF+E$ (the tilde denotes part of a matrix - the first p column). It can be shown, as mentioned above, that this solution has a least square property: among all approximation of rank p of X, it minimizes the Frobenius norm of E, $\|E\|_F$.

Thus, we can define the unweighted factorization of rank *p* of X as:

$$\{G,F\}=\arg\cdot\min_{G,F}\|X-GF\|_F \tag{2.3.3}$$

where G and F are required to be of previously selected rank *p*. The solution is optimal (i.e. minimum variance) if and only if all the variable X_{ij} to be fitted are of the same accuracy (Paatero and Tapper, 1994).

2.3.1 Scaling of the data matrix

The singular value decomposition of a matrix X is not invariant with respect to scale changes of columns or rows of X, i.e. if the units of measurement are changed from one row or column to another, the SVD gives rise to different matrix decompositions. In order to solve this problem, various scaled forms of X in PCA have been used. Subsequently, these scaling transformations have been studied in connection to a weighted least square, where error estimates are available for the measured data (Paatero and Tapper, 1994; Paatero and Tapper, 1993).

Using the PCS notation, given the input matrix X and the matrix Y = GF that define X, a weighted least square fit of X is performede by the following minimization expression:

$$\{G,F\}=\arg\min_{G,F}\sum_{ij}w_{ij}\left(x_{ij}-y_{ij}\right)^2 \tag{2.3.4}$$

where w_{ij} are the weight corresponding to each x_{ij}. Varying w_{ij}, the best solution is obtained when each scaled factor w_{ij} is equal to $\left(\text{std.dev}\left(x_{ij}\right)\right)^{-2}$, that is the inverse of the squared standard deviation of $x_{ij.}$

The most general scaled form of PCA is defined with the help of diagonal matrices. Applying the SVD the solution can only be optimal if the standard deviation matrix, σ, is of rank one. If, however, rank(σ)>1, the optimality condition cannot be fulfilled and SVD cannot be an optimal least square (LS) method. Among the different type of scaling, the standard scaling are the column norm scaling and the row norm scaling, i.e. the accuracy of all elements on any single column/row are assumed to be equal. They are not recommended as a general purpose tool in physical or chemical application, because the scalings are limited to treating whole columns or rows, while individual treatment of matrix elements is impossible (Paatero and Tapper, 1993). Alternative scaling deals with the possibility of find a rank one matrix which approximates σ and use this matrix to obtain the best possible scaling (Paatero and Tapper, 1993).

2.3.2 Rotations

Any non-singular square matrix T defines a rotation of the solution by:

$$X = GF + E = GTT^{-1}F + E = \overline{G}\overline{F} + E \qquad (2.3.5)$$

where the new rotated factors are $\overline{G} = GT$ and $\overline{F} = T^{-1}F$.

A rotation does not affect the residual matrix E. All rotation can be represented as sequence of the so-called *elemental rotation*, expressed by the following pair of matrices (Paatero and Tapper, 1994):

$$T = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}, \qquad T^{-1} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} \qquad (2.3.6)$$

These matrices represent the addition of one column of G, multiplied by a scalar factor *a*, to another column, and subtraction of one row of F, multiplied by the same scalar facto *a*, from another row.

Rotations are used to eliminate negative elements from factors G and F, but there is however the risk to produce new negative elements, because increasing the values on one side simultaneously decreases the value on the other side (see elemental rotation coefficients *a* and *–a*)

The introduction of elementary rotation leads to the definition *p-rotatable factorization* (p means '*positive*') (Paatero and Tapper, 1994):

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

2.4 *Positive Matrix Factorization*

Positive Matrix Factorization (PMF) is a recent type of receptor model, developed by Dr. Pentti Paatero (Department of Physics, University of Helsinki) in the middle of the 1990s (Paatero and Tapper, 1994; Anttila *et al*., 1995), in order to develop a new method for the analysis of multivariate data that resolved some limitations of the PCA. One of the main positive aspects is the use of know experimental uncertainties as input data which allow individual treatment of matrix elements. This becomes increasingly important with the introduction of the Guide for Expression of Measurements (GUM) and the derived Guide for Quantification of Analytical Measurements (QUAM), which are nowadays commonly accepted references underlying numerous national and international standards (ISO/IEC, 2008; Ellison *et al.*, 2000).

However, point-by-point scaling results in a scaled data matrix that cannot be reproduced by a conventional factor analysis based on the SVD (Paatero and Tapper, 1993).

Positive Matrix Factorization is as a weighted factorization problem with non-negativity constraints which, given the matrices X (input data matrix) and σ (uncertainties data matrix) and a selected rank *p*, is defined in the 2-dimensional case by the following expressions:

$$X = GF + E, \quad G: n \times p, \quad F: p \times m \tag{2.4.1}$$

$$G_{ik} \geq 0, \qquad F_{ik} \geq 0 \tag{2.4.2}$$

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} E_{ij}^2 / \sigma_{ij}^2 \tag{2.4.3}$$

$$\{G, F\} = \arg\min_{G,F} Q \tag{2.4.4}$$

The problem is symmetric with respect to the rows and column of the matrix X and the factors G and F: this is a '*bilinear model*'.

Its resolution is a difficult task, because it has two different non-linearities: inequalities and products of unknowns.

Two are the main algorithms used to solve this problem: PMF2 and ME-2 (Multilinear Engine), discussed in the following.

The introduction of the standard deviation matrix σ into the model creates a link between the factor model and the physical reality. Also, handling the standard deviation of certain non representative data avoid their discarded or the formation of a noise that arise in PCA.

A more detailed PMF description and its mathematical algorithm are described in Chapter 3.

2.5 *Comparison between PMF and PCA*

The main difference between PCA and PMF is that in the fist one the solution forms a hierarchy and so, a higher dimension (a higher number of factors) contains all the factors of the lower dimension, while in the last one the factors are not orthogonal and so there is no hierarchy. However, when rotation is applied to PCA, the factors are not anymore orthogonal.

Usually in physical sciences, the factors have not the orthogonality property so its missing in PMF is not significant. Moreover, PMF produces non-negative distributions (factors) by definition and this aspect precludes the orthogonality.

Resolving PMF algorithms is however slower than PCA but, on the other hand PCA is simpler to use because of less parameters to control.

Other different aspects between these two methods concern the rank of the standard deviation matrix and the p-rotatable property of SVD. In fact, different cases can be presented, as summarized hereafter (Paatero and Tapper, 1994):

1. The matrix σ is of rank one and SVD of X is p-rotatable: with PCA the matrix can be scaled correctly and factorization by SVD is optimal. The factorization by PMF is always optimal, because it always uses 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 and Q(PMF)=Q(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.

2. The matrix σ is of rank one but SVD of X is not p-rotatable: it's impossible to rotate the SVD-derived factorization, while PMF will produce the desired solution. PMF solves the problem, PCA does not.
3. Rank(σ) > 1: correct scaling is not possible with SVD; it's only possible to approximate the σ with a matrix of rank one, leading to loss of information. 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.

From this it becomes apparent that in conclusion, PMF is generally more powerful than the best possible PCA or at least equivalent to PCA. In exchange, PCA is that computing a SVD is much faster than in case of using PMF.

3 Positive Matrix Factorization

3.1 *Introduction to PMF*

In the previous chapter a new receptor model has been introduced, namely Positive Matrix Factorization (PMF). At the beginning PMF has been used in air pollution and source apportionment studies (Polissar *et al.*, 1999; Lee *et al.*, 1999) and in precipitation study (Juntto and Paatero, 1994; Anttila *et al.*, 1994). Also recently, applications on air quality and source apportionment (Xie and Berkowitz, 2006; Begum *et al.*, 2004) have been carried out. In addiction, in the latest years, PMF has been applied to lakes sediments (Bzdusek *et al.*, 2006), wastewater (Soonthornnonda and Christensen, 2008; Singh *et al.*, 2006) and soils (Vaccaro *et al.*, 2007; Lu *et al.*, 2008).

As mentioned above, Positive Matrix Factorization differs from the customary factor analysis models such as PCA by the property to take into account standard deviations of observed data values and to introduce the constraint of non negativity (hence the use of the term "*positive*") of all the factor matrices G and F elements in order to have physically meaningful solutions. It is thus a weighted least square problem in which a certain number of factors have to be determined in order to minimize an '*object function*'. As stated above, this problem cannot be solved by SVD. The input data are a multivariate data set containing the measured data and the corresponding uncertainties data matrix.

One of the main features of PMF results is their quantitative nature: it is possible to obtain the composition of the sources determined by the model. In contrast, the results of PCA are qualitative as they can only distinguish variables that tend to appear together from those ones that do not (Paatero, 2004).

Moreover, in contrast to customary factor analysis models, PMF model has been implemented to handle non representative data such as "*below detection limit*", missing data and outliers. This is an important property as it prevents the rejection of such values and hence the reduction of the initial data set. These and other positive aspects are detailed described in the following section.

Different approaches to resolve PMF algorithm have been studied, both for usual 2-dimensional matrices and 3-way arrays. The firsts programs developed by Paatero are called respectively PMF2 and PMF3 (Paatero, 2004a; Paatero 2004b) and later on the algorithm has been extender to arbitrary multilinear models by means of the program *Multilinear Engine* (ME), (Paatero, 1999). In the latest years other resolving techniques have been developed, starting from Paatero's PMF equations, like a new PMF formulation by Bzdusek (Bzdusek *et al.*, 2006). Moreover, given the importance of receptor models in scientific research, the United States Environmental Protection Agency (US-EPA) has developed a standalone version of PMF, EPA PMF 3.0, freely distributed (Norris *et al.*, 2008). EPA PMF 3.0 is based on ME-2 (ME second version; Paatero, 2007)

To have a clear distinction between PMF as a model and the name of the programs, the model is designated as **PMF** while the programs used to solve the model are designated **PMF2**, **PMF3** and **ME-2**.

3.2 *PMF2 algorithm*

PMF2 is used to solve 2-dimensional problems by means of the following bilinear model:

$$X = GF + E$$

or, in component form:

$$x_{ij} = \sum_{p=1}^{p} g_{ip} f_{pj} + e_{ij} \qquad i = 1\ldots m;\ j = 1\ldots n \qquad (3.2.1)$$

where X is the measured data matrix, G and F are the matrices to be determined and E is the residual matrix (the unexplained part of X). Due to the factor linearity, the matrices G and F can be exchanged without changes in the matrix X. In a practical example X can be viewed as a matrix contained measured value of certain variable, G the contributions matrix of the identified sources and F the matrix characterizing each sources. The elements of G and F are constrained to assume positive value and this corresponds to the idea that no sources may emit negative amounts of physical substances.

The expression of the object function to be minimized as a function of G and F is given by:

$$Q(E) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 \qquad (3.2.2)$$

where σ_{ij} are the known uncertainties for each data value x_{ij}, so that the optimum weight in the least square fit are 1/ σ_{ij}. In this way the PMF problem is then identified as a minimization of Q(E) with respect to G and F, and under the constraint that each elements of the matrices G and F is to be non-negative.

As in PCA, also in PMF there is rotational ambiguity. Starting from the T definition given in Section 2.3.2 and the identity:

$$GF = GTT^{-1}F \qquad (3.2.3)$$

the expressions GT and $T^{-1}F$ represent a pair of factors which are '*equally good*' (same goodness of fit) as the original pair, G and F.

The original basic algorithm used to determine the 2-dimensionl solution was the so-called Alternating Regression (AR): one of the matrices G and F is taken as known and the object function Q is minimized with respect to the other matrix; then their roles are interchanged. This process is continued until convergence (Paatero and Tapper, 1993). However, this process can be slow if the factors are far from being orthogonal (needing up to thousand of steps).

In order to improve the model performance the AR algorithm was modified by computing G and F steps where both these matrices are changing simultaneously. Starting for example from $G = G_0$ and $F = F_0$ the iteration consists of the repetition of the following three basis steps:

- minimize for $G = G_0 + \Delta G$ while keeping $F = F_0$ constant;
- minimize for $F = F_0 + \Delta F$ while keeping $G = G_0 + \Delta G$ constant;
- minimize for the extension coefficient α in

  $$(G_0 + \alpha\Delta G)(F_0 + \alpha\Delta F) = X + E$$

  where ΔG and ΔF are as determined from the first and second steps.

This algorithm is fast and typically the convergence needs 30 to 100 steps.

Subsequently, as a generalization of the latter AR algorithm, PMF2 algorithm was created by Paatero. It is able to simultaneous vary the elements of G and F in each iterative steps and have a faster convergence.

3.2.1 Mathematical algorithm

In PMF2 the object function Q(E) assumed a more complicate formula than the simpler Equation 3.2.2 one, because the addition of a correct implementation of the non-negativity constraint and of two terms to reduce the rotational ambiguity. The objective is to minimize the expression $Q(G_0 + \Delta G, F_0 + \Delta F)$ with respect to ΔG and ΔF simultaneously. Starting from arbitrary matrices ΔG and ΔF in the factor spaces G and F of dimension nxp and mxp respectively, during each such '*full*' step of the iteration there would be (m+n)p unknowns to solve. Because the number of unknowns in this 2-way models may be very large, the full model is not very efficient with respect the computational workload. Thus, PMF2 works on two restricted spaces (for computational details see Paatero, 1997).

Here we summarize only the main steps of the PMF2 algorithm with a detailed description to be found in the above mentioned article.

The new object function, called *enhanced object function* is defined as:

$$\begin{aligned}\overline{Q}(E,G,F) &= Q(E) + P(G) + P(F) + R(G) + R(F)\\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 - \alpha\sum_{i=1}^{m}\sum_{k=1}^{p}\log g_{ik} - \beta\sum_{k=1}^{p}\sum_{j=1}^{n}\log f_{kj} \qquad (3.2.4)\\ &+ \gamma\sum_{i=1}^{m}\sum_{k=1}^{p} g_{ik}^2 + \delta\sum_{k=1}^{p}\sum_{j=1}^{n} f_{kj}^2\end{aligned}$$

P(G) and P(F) are called *penalty function* and prevent the elements of the factor matrices G and F from became negative. R(G) and R(F), called *regularization function*, are used to remove some rotational indeterminacy and to control the scaling of the factors. The coefficients α, β, γ and δ of the Q equation control the strength of their respective. For efficiency reasons the log function of the penalty term was approximated by a Taylor series expansion up to quadratic terms (Paatero, 1997).

To solve each 'full' step the algorithm use the Gauss-Newton and Newton-Raphson numerical method and the Cholesky decomposition. Between these steps, rotational substeps are performed: the algorithm determines a rotation T and its inverse T^{-1} so that the new factor matrices GT and $T^{-1}F$ minimize the enhanced object function. Taking into account the definition of rotation, the main part of $\overline{Q}$ does not change because the matrix T does not change the residual matrix and so the minimization is only related to the penalty and regularization terms. These rotations lead to a fast computation.

Rotational ambiguity

As above discussed, PMF have a rotational ambiguity, which make the solution of the algorithm not unique. To illustrate this problem we can consider the example of two measured elements, iron ad silicon, in environmental samples, illustrated in the following picture (Paatero *et al.*, 2002)

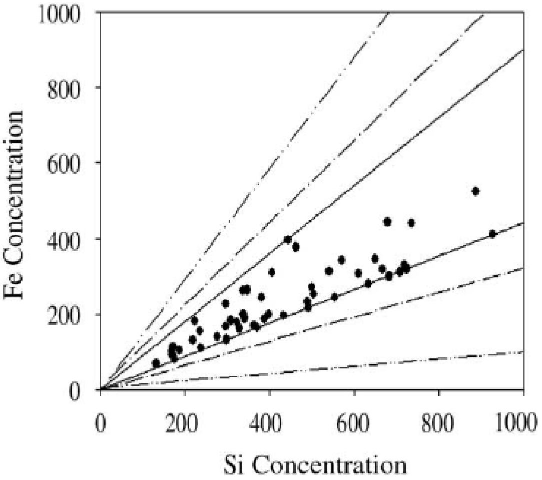


Figure 3.1 – Example of rotational ambiguity (Paatero *et al.*, 2002)

In order to reproduce the sources profile we need two factors, as the two parameters are not correlated each other; however we can choose one of the many profile which range from the Cartesian axes to the solid lines. In conclusion, without additional information the source profile cannot be completely determined.

Some of the rotational ambiguity is however removed by the non-negativity constraint of the matrices G and F as the transformation described by Equation 3.2.3 is only acceptable if it does not produce negative elements in these matrices. In the extreme case in which all the rotations are forbidden the solution is unique. Actually there are different possible rotations so the issue is to determine the optimal solution that better represents the problem under analysis

But what happened when an elementary rotation with a *r* coefficient is applied?

If the *r* coefficient is positive, by the effect of the matrix T (see Section 2.3.2) the G elements are spread towards positive value meanwhile F elements are spread toward negative values. On the contrary, if *r* has a negative value the elements of the matrices G and F are changed vice-versa. In this manner, if all the matrices elements are positive, all the rotations are possible (Paatero *et al.*, 2002).

However, as can be proofed by the multivariate statistic analysis theory, if a sufficient number of G (or F) elements assumed a priori zero value, then there is not rotational ambiguity and the solution is unique.

In PMF2 algorithm rotations are implemented in the iterative steps by means of the so-called *FPEAK* parameter, ϕ, which can assume positive or negative value (the zero-value correspond to a free rotation solution also called centred solution). If the factor matrices have all non-zero values there is not a logical common sense in selecting a specific ϕ value, but the entire ϕ domain can be explored. Also in PMF2 the rotations are controlled by means the T matrix but an additional control is provided by the regularization terms in the enhanced object function (Equation 3.2.4). This latter is also used in PMF3 and ME-2 algorithm where there is not a T-corresponding mechanism for controlling rotations.

3.2.2 Errors

If the error matrix is not known (i.e. analytical uncertainties are missing) an estimation of the data uncertainties could be performed using known concentrations (or variables amounts) and the limit of detection values.

In a recent source apportionment study based on PM2.5 (Particulate Matter with an aerodynamic diameter of 2.5 μm or less) data (Kim and Hopke, 2007), PMF results from known uncertainties and estimate uncertainties have been compared. The resulting factors are similar and little difference seems to be due by differences in true uncertainties from the different laboratories.

In literature different kinds of uncertainties estimation are reported. As an example, in the study of daily precipitation data, described in Juntto and Paatero, 1994, the standard deviation associated to a single measurement of ions *j* is calculated by:

$$\sigma = e_j + d_j x$$

where x is the ion *j* concentration. The values e_j and d_j are determined experimentally for each ion in order to include in the calculated standard deviations about the 80-90%

of the data. Like this, for small concentrations (near the limit of detection) the e_j value prevail on the standard deviation, while for large concentrations became more important the d_j x member.

The parameters used in the error estimates (in the above example: ej and dj) could be determined by trial and error, varying their values until they produce the best fit. In example, in Reinikainen *et al.*, 20001, the parameters are adjusted in order to obtain approximately equal scaled residuals for all the data. Instead, in Polissar *et al.*, 2001, the parameters corresponding to the best fit are evaluated analyzing Q values, scatterplots, distributions of the residuals and results from multiple regressions.

It is important to note that when an input data-set contains below detection limit data and/or missing data, referred as non representative data, they can also be handle by the model, avoiding loss of information. However, in this case their errors and concentrations have to be estimated. This can be done in different ways and in the following section some error and data estimates are listed for dataset including this type of data.

Error model (EM)

Alternatively, when the uncertainties of the input data x_{ij} are not known, PMF2 can compute error estimates for x_{ij}.(Paatero, 2004). This is done by means of the three codes C1, C2 and C3, the *Errormodel* (EM) and the three arrays T, U and V. As explained in the following, the Errormodel makes a choice of the equation used to determine the standard deviation matrix, named next by S.

In the simplest case in which all the X elements have either the same accuracy or the same relative error, only the C1 and C3 value have to be set, corresponding to the matrices T and V respectively. Usually, the U matrix (and so the C2 value) is not used unless in the case of Poissonian situations.

The S matrix is computed by two ways: in the first one this matrix is determined before the iterative steps are started (EM = –12); in the latter, S is determined during each step using the fitted value y_{ij} in place of the x_{ij} values (EM = –10, –11, –13, –14).

Following a description of the different error models:

- EM = –12. The equation used to determine the s_{ij} value is given by:

$$s_{ij} = t_{ij} + u_{ij}\sqrt{|x_{ij}|} + v_{ij}|x_{ij}|$$

The T matrix corresponds to the x_{ij} uncertainties matrix and the V matrix to the relative errors one. Alternatively, if the T, U and V matrices contain each one the same elements, then their values can be replaced with the corresponding C1, C2 and C3 values. Typically C1 is chosen equal to the detection limit and C3 in the range 0.01–0.1.

- EM = –10. In this error structure it is assumed that data and uncertainties have a lognormal distribution. Assuming also there is a measurement error with standard deviation equal to t_{ij} and be v_{ij} geometric standard deviation logarithm, the S matrix is iteratively calculated by:

$$s_{ij} = \sqrt{t_{ij}^2 + 0.5v_{ij}^2 \left|y_{ij}\right|\left(\left|y_{ij}\right| + \left|x_{ij}\right|\right)}$$

  For this model the fitted Q value is greater than the expected one, corresponding to the problem degree of freedom.

- EM = –11. It is assumed a poissonian data distribution with μ_{ij} parameter equal to GF. During the iteration steps s_{ij} values are given by:

$$s_{ij} = \sqrt{\max\left(\left|\mu_{ij}\right|, 0.1\right)}$$

- EM = –13. Like the EM = –12 structure the s_{ij} values are given by the same formula, but now they are computed at each iterative step replacing the x_{ij} values with the relative fitted y_{ij} values.

- EM = –14. The s_{ij} value are computer by means the following equation:

$$s_{ij} = t_{ij} + u_{ij}\sqrt{\max\left(\left|x_{ij}\right|, \left|y_{ij}\right|\right)} + v_{ij}\max\left(\left|x_{ij}\right|, \left|y_{ij}\right|\right)$$

  This option is recommended in environmental work as an alternative method to the EM = –12, although the processing time is greater.

When the uncertainties matrix is read from an external file (i.e. the matrix is computed by the user as one of the error estimates method described in section 2.2.3), only the T matrix is read in the .INI file, so C2 = C3 = 0 and EM = –12.

3.2.3 Non-representative Data

Below detection limit and missing data

Typically, elemental concentrations data set or physical-chemical parameters data set can contain below-detection-limit values (BDL) and/or missing values (MV). In the below-detection-limit data the values are below the method detection limit so we only know they are small. In the case of missing data the values could not be determined and hence they are totally unknown. In order to avoid rejection of this data or, in the worst case, rejection of all the variables related to the same sample (input matrix must not contain null values) as done in PCA, PMF is able to handle BDL ad MV by means of different type of data values estimates and associates error estimates. Also, when the data uncertainties are not known it is possible to associate calculated error estimates.

In literature different types of data and errors estimates can be found and here below some example are reported.

Polissar (Polissar *et al.*, 1998) in a PMF study for atmospheric samples has suggested the following estimates:

$x_{ij} = \nu_{ij}$ $\quad\sigma_{ij} = u_{ij} + DL_{ij}/3$ $\quad$ for determined values ν_{ij}

$x_{ij} = DL_{ij}/2$ $\quad\sigma_{ij} = \overline{DL}_{ij}/2 + DL_{ij}/3$ $\quad$ for below detection limit values

$x_{ij} = \overline{v}_{ij}$ $\quad\sigma_{ij} = 4 \cdot \overline{v}_{ij}$ $\quad$ for missing value

where u_{ij}, DL_{ij} and $\overline{v}_{ij}$ are the analytical uncertainty, the method detection limit and the geometric mean of the measured concentrations respectively, for sample i and parameter j.

In this case, the detection limit specified the error estimates for low data value while the uncertainties provided the estimation of errors for high data values. According to this equations, relative error estimates for below detection limits value range from 100% to 250%, while for missing value are equal to 400%.

Xie and Berkowitz, 2006, in an application to hydrocarbon emissions, use the previous estimates introducing the additional percentage parameter C2:

$x_{ij} = \nu_{ij}$ $\quad\sigma_{ij} = MDL_{ij}/3 + C2 \cdot x_{ij}$ $\quad$ for determined values ν_{ij}

$x_{ij} = MDL_{ij}/2$ $\quad\sigma_{ij} = MDL_{ij}/2 + MDL_{ij}/3$ $\quad$ for data below MDL_{ij}

$x_{ij} = \overline{v}_{ij}$ $\quad\sigma_{ij} = 4 \cdot \overline{v}_{ij}$ $\quad$ for missing value

where C2 is the percentage parameter determined by trial and error and MDL stand for Minimun Detection Limit.

In an atmospheric aerosol study Polissar *et al.*, 2001, replace below detection limit data with half detection limit values and missing data with mean concentration. For determined values the errors formula used is:

$$\sigma_{ij} = \sqrt{a_j u_{ij}^2 + b_j DL_{ij}^2}$$

where u_{ij} are the analytical uncertainties and DL_{ij} are the analytical detection limits for sample i and element j. The variables a_j and b_j are the scaling factor for the weight associated to each element and their values are chosen by trial and error.

For below-detection-limit data, the term u_{ij} was set equal to zero, so that only the second term of the error equation is used. For missing values, the error estimate equal 25 times the mean element concentration

PMF2 allows an automatic handling of missing value by the use of the optional parameters "*Missingneg* r", with *r* as a decimal value, and "*BDLneg* r1 r2", with *r1*<0. For detailed information see Paatero, 2004a. However these options must be used with caution.

Outliers

Outliers are extreme values that differ from the mean trend of all the data. They can occur for various reasons, for instance because of a sample contamination that affect all the elements in a row of the data matrix (a sample) or a laboratory error that affect only one element of the data matrix. Such very high or very low value can also be true outliers, but in either case they have a significant influence on the solution. Thus, PMF offer a so-called "*robust mode*" where the outliers influence is reduced. This robust factorization is a technique of iterative reweighing of the individual data values based on the *Huber influence function*, that modify the object function Q.

In the non-robust formulation we can express Q in terms of the scaled residues $r_{ij} = e_{ij}/\sigma_{ij}$

$$Q(r) = \sum_{i=1}^{m}\sum_{j=1}^{n} Q(r_{ij}) \qquad \text{where} \qquad Q(r_{ij}) = r_{ij}^2 \tag{3.2.5}$$

Usually, the influence function $\psi(r)$ is defined as half of the derivative of the functional Q(r) (Paatero, 1997):

$$\psi(r_{ij}) = 0.5\frac{\partial}{\partial r_{ij}}Q(r_{ij}) = r_{ij}$$

For each x_{ij}, this equation indicates that values with high residue have a high influence, but this is not the correct way to handle outliers because they represent data of a poorer quality.

Robustness is hence achieved by constructing an appropriate influence function, the Hubert function, which limits the maximum strength that each data can bring to the fit. It is defined by:

$$\psi^H(r_{ij}) = \begin{cases} -\alpha & \text{if } r_{ij} < -\alpha \\ r_{ij} & \text{if } -\alpha \le r_{ij} \le \alpha \\ +\alpha & \text{if } r_{ij} > \alpha \end{cases}$$

where α is the outlier distance (the distance for classifying the observation as outliers).

The object function corresponding to ψ^H is denoted by Q^H and the least square formulation becomes:

$$Q^H(E) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{e_{ij}}{h_{ij}\sigma_{ij}}\right)^2 \qquad h_{ij}^2 = \begin{cases} 1 & \text{if} \left|e_{ij}/\sigma_{ij}\right| \le \alpha \\ \left|e_{ij}/\sigma_{ij}\right|/\alpha & \text{otherwise} \end{cases}$$

In this manner, the outliers are not rejected but they are handled as they stay at the distance $\alpha\sigma_{ij}$ from the fitted value. This method however is not applied to outlier with very low values respect the mean observations.

High noise variables

In environmental studies it may happens either that some variables present a higher noise than others or the noise is greater than the signal (in some situation the signal may also be absent). So the problem is to identify and handle these variables.

In Paatero and Hopke, 2003, they list the variables using the ratio signal to noise (S/N): *weak* variable contain signal and noise in similar quantities, while *bad* variables contains much more noise than signal. In order to have a general numerical definition base on this ratio Pattero and Hopke define a variable is weak if:

$$0.2 < \frac{S}{N} < 2$$

Like this a variable is bad if S/N<0.2.

If the detection limits (DL) are known and the below detection limit values are replaced with DL/2, then the relation defining a weak variable is the following:

$$0.2 < \frac{\sum_{\{i|x_{ij}>\delta_{ij}|\}} x_{ij}}{\delta_j n_{DLj}} < 2$$

where n_{DLj} is the number of below detection limit value in column j and δ_j is the mean detection limit in column j.

Since to each variables is associated a corresponding weight (inverse of the errors), one must may pay attention to do not have either downweighting (weight too high) or overweighting (weight too low). Relating to weak variables it is recommended to downweight them by a factor of 2 or 3 in order to be sure that their noise does not affect significantly the result. Bad variables can be rejected from the analysis or, if they are kept in the dataset, it is recommended to downweight them by a factor 5 to 10 (Paatero and Hopke, 2003)

3.2.4 Explained variation

The Explained Variation (EV) is a dimensionless quantity describing how much each computed factor explained a row (EV of G) or a column (EV of F) of the input data matrix, X. The explained variation matrix is defined by Paatero, 2004b, considering that the X values are explained by both the p factors and the residuals. In this way, residuals form a *(p+1)* additional factor also called NEV (Not Explained Variation) as they represent the X part not explained by the p factors. The EV elements values range from 0.0 to 1.0 corresponding to no explanation and complete explanation respectively.

In the G matrix case, EVG is a *nx(p+1)* matrix where the *(p+1st)* column correspond to NEVG; their elements are give by the equations:

$$EVG_{ik} = \frac{\sum_{j=1}^{m} |g_{ik} f_{kj}| / s_{ij}}{\sum_{j=1}^{m} \left(\sum_{h=1}^{p} |g_{ih} f_{hj}| + |e_{ij}| \right) / s_{ij}} \qquad \text{for } k = 1, \ldots, p$$

$$\mathrm{NEVG}_{ik} = \frac{\sum_{j=1}^{m} \left|e_{ij}\right| \Big/ s_{ij}}{\sum_{j=1}^{m} \left(\sum_{h=1}^{p} \left|g_{ih} f_{hj}\right| + \left|e_{ij}\right| \right) \Big/ s_{ij}} \qquad \text{for } k = p + 1$$

The first equation gives information of how much each factor (1, …, p) explains the ith row of X; in example, in the case of a data set made by measurement of *j* parameters in *i* samples, EVG$_{ik}$ describe the amount of ith sample explained by the kth factor.

Instead, the second equation describes how much the residues explain the ith row of X or, like the above example, describes the amount of ith sample not explained by the p factors. By definition, the EVG and NEVG sum must be equal to one.

Similar equations are used to determine the EVF matrix, where the sum over j is now substituted with an '*i*' sum. This is a *(p+1)xm* matrix where the last row indicate the NEV of F that is the amount of the jth variable unexplained by the p factor (see above example). However, it is a practical rule to consider unexplained a variable when its NEVF value exceeds 0.25. In this case it is advisable to decrease the weight of the variable such as their residues are approximately between – 1 and +1. In source apportionment cases the explained variation of F is used in order to qualify the sources since a factor explaining a large amount of one or more parameters can be identify according to the origin of those parameters.

The information carried out by the explained variation must be handled with care, mainly in the presence of high outlier values in the data matrix because of the explained variation are computer using the original standard deviations. For more realistic value it is opportune to manually decrease the weigh of outliers.

4 PMF – Toolbox

In this chapter we want to focus the attention on the choice of the optimal solution. As described in the Appendix, there are many parameters involved in the determination of the factor matrices and, change one of them, may lead to a different solution. This is why it is not correct to investigate only one parameters combination, while several runs with different conditions should be made.

To better explain how to manage the different PMF parameters to find the optimal solution, an example based on an existing data-set is reported. This data-set is referred to XRF (X-Ray Fluorescence) analyses on sediments, extracted from 12 alpine Italians lakes; for each lake from 17 to 20 samples had been collected.

4.1 *Initialization*

As to the first, in order to produce different solutions to compare each other, we make different runs with the FPEAK parameter equal to zero (no rotations), changing at every turn the number of factors. Customary the starting number of factors is 2 and the maximum number may be of the order of tens; even if the number of factor seems to be small or large in relation to the problem under analysis, we should considered these solutions until we are sure of their wrongness.

The next step is to investigate different values of the FPEAK parameter for each number of factors. However this may lead to a high number of analyses and so it should be more easily to reject some numbers of factors, using the tools described in the next section, in order to reduce the number of possible solutions.

FPEAK can assume positive and negative values, but different studies reported that too high positive and negative values lead to a poorer fit. So the correct values to be investigated vary from -2 to 2 with a step chosen by the user (ideally a 0.2 step is used as an intermediate choice).

4.2 *Determining the number of factors*

Obviously, the solution is dependent from the number of factors and it is necessary to retain the one that optimally describe the problem under analysis. However the solution has also a rotational ambiguity so it is better to analyze together these two important parameters.

In this section we describe the methods used to select the correct number of factors starting from the central solutions (FPEAK=0); the selection involves the analysis of output parameters computed during the fit.

4.2.1 Analysis of Q value

Before describing the procedure used to select a range of number of factors starting from the Q values computed by the fit, we have to look at the following guideline, based on the expected Q value.

In weighted-least-square problems if the uncertainties associated to the variables are correct, the Q function should be distributed as a chi-square, χ^2, distribution (Q value is a χ^2 value). The degrees of freedom are an important parameter of this distribution as they correspond to the χ^2 expected mean value; this number is calculated by subtracting the number of bond (or free parameter) from the number of data points. In the two-dimensional approach the free parameters of the matrices product GF is given by *(n + m)xp*; if also the rotational ambiguity is to be considered then the introduction of the matrix T *(pxp)* makes the free parameters equal to *(n+m–p)xp*. Considering the Q expression, the resulting degrees of freedom are $\upsilon = nxm - (n + m - p)xp = (n - p)x(m - p)$ (Paatero and Tapper, 1993) and consequently the expected Q is given by:

$$Q^{exp} = (n - p)x(m - p)$$

If the data matrix is expected to be very large, in example $mxn >> p(m + n)$ then $Q^{exp} \approx$ *mxn*, that is the expected value of Q may be approximated to the number of data points.

Now it is clear that Q^{exp} value gives important information about the quality of the fit since the optimal solution should have a Q not too different from the Q^{exp} value. This may be a tool for selecting the optimal number of factors (Bzdusek *et al.*, 2006) but in some cases, the optimal solution does not satisfy this requirement, in example when a dataset contains much weak variables or the uncertainties are not well known. In this latter case, the standard deviations can be '*settled*' so that the expected Q is close to the theoretical one (see also Section 3.2.2 about choosing errors estimates that improve the quality of the fit).

Using the data-set of alpine lakes, from 2 to 8 number of factors have been tested (with FPEAK=0) and the resulting Q/Q^{exp} values are plotted in the following figure:

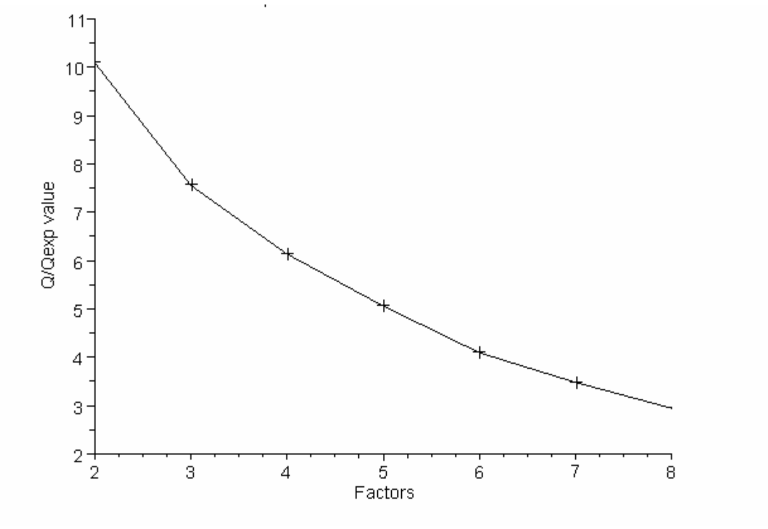


Figure 4.1 – Q/Q^{exp} values from different number of factors

The Q/Q^{exp} value steadily decreases except from factor 2 to 3 where the slope is greater; this suggests rejecting the solution with 2 factors. Q ranges from 3 to 10 times the Q^{exp} value and this is could be due to the uncertainties associated with the data that are not known, but are computed with the formula used by Xie and Berkowitz, 2006 (see also Section 3.2.3).

Also, for each number of factors, we computed different initial runs to assess Q stability, using different starting points (see Appendix, .INI file and Multiple results). Local minima occurred in results from 6, 7 and 8 number of factors suggesting that too many factors are used (see Appendix, Multiple results).

4.2.2 Analysis of residuals

Another method is based on the analysis of the scaled residuals of the model (Juntto and Paatero, 1994). The scaled residuals may in fact be used in order to detect data anomalies. If the input data and the model are correct, the plot of scaled residual values against their occurrence shows a random distribution with no positive or negative divergences. Customary the majority of them is located from -2 to +2 (Juntto e Paatero, 2004).

On the other hand if they fluctuate outside of this range it is possible that either the chosen number of factors is not correct or there is some noise in the variables (i.e: outliers or downweight).

Moreover, it may be happened that the scaled residuals are large for certain variables, because the associated standard deviations are set to too small values, so it would be better to increase their values. Opposite to this, if the scaled residuals are too small, the standard deviation may be set too large or the variable is explained by a unique factor. This latter case may occur naturally but such situation may also occur when high standard deviations have been specified for a noisy variable (Paatero, 2004a).

However, it is necessary to pay attention at these approaches as it may be happens that a factor have a good fit even if it is not "interpretable"; in example it explain only one variable (Huang *et al.*, 1999).

In the example of alpine lakes, after a first step of analysis of residuals, data uncertainties have been modified accordingly to preceding residuals plots, increasing the errors for those element that show a bad residuals distribution (Free *et al.*). In **Error! Reference source not found.** the plot of Mn residuals is reported, before and after the associate uncertainties have been increased by a factor 1.5.

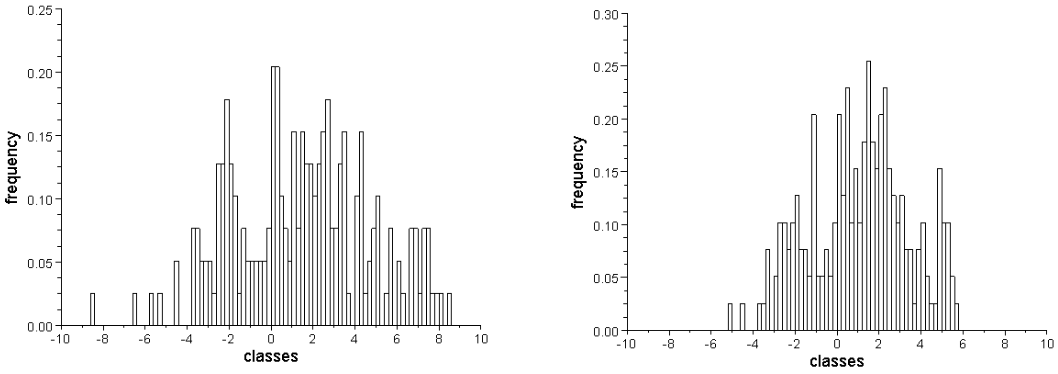


Figure 4.2 – Residual plots for Mn. Left: before increase Mn uncertainties. Right: after increase Mn uncertainties

As it can be seen, increasing the uncertainties lead to a residuals distribution closer to zero and hence to a better fit. However residuals are still not well located between -2 and +2 and this is probably due to the variability of the elements concentrations. In fact, as the geochemical data are referred to different lakes, the possible presence of different lithologies may causes extremely inhomogeneous elements concentrations, as in the case of Pb, illustrated in Figure 4.3.

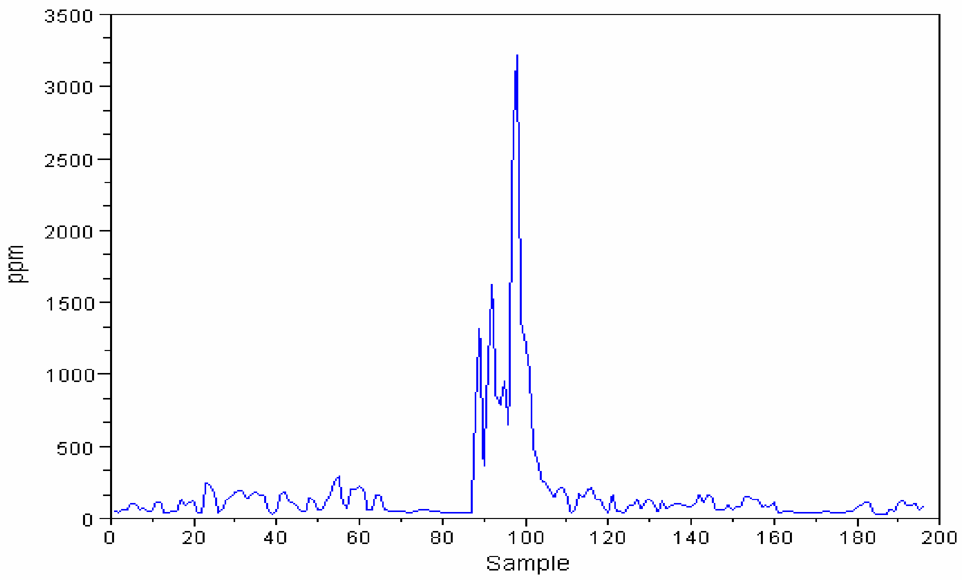


Figure 4.3 – Spatial plot of Pb

The points where the concentrations are very high are all referred to one of the alpine lakes; hence these data are not outliers because they correspond to real concentrations. This situation could lead to some ambiguity in the analysis of results, in fact in the plot of Pb residuals (Figure 4.4) the distribution is bimodal and one could deduce a poorer fit. Actually this bimodal distribution reflects the original spatial distribution (see also Polissar *et al.*, 1998).

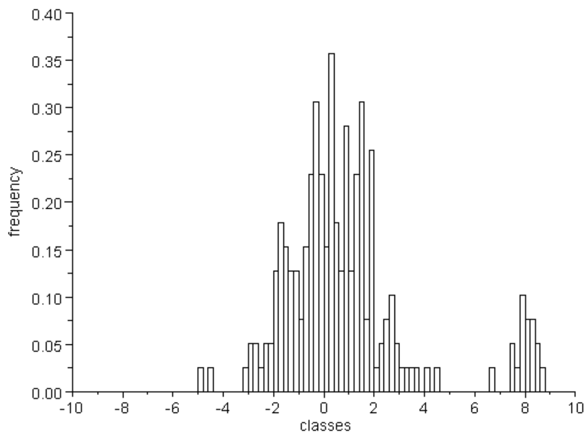


Figure 4.4 – Plot of Pb residuals

4.2.3 IM and IS

In order to reduce the range of the meaningful number of factors, two parameters named IM and IS have been used in Lee *et al.*, 1999. Starting from the scaled residual matrix R, these parameters are computed as follow:

$$\mathrm{IM} = \max_{j=1\ldots m}\left(\frac{1}{n}\sum_{i=1}^{n} r_{ij}\right)$$

$$\mathrm{IS} = \max_{j=1\ldots m}\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(r_{ij} - \bar{r}_j\right)^2}\right)$$

where $\bar{r}_j$ is the mean over the ith row.

As stated from this expression, IM represents the jth variable with greater scaled residuals mean and so the less accurate one. Instead IS reproduces the jth variable with greater scaled residual standard deviation and so the more imprecise fit.

Plotting these parameters against the number of factors it is possible to reject some of them from further analysis as IM and IS show a drastic decrease when the number of factors increase up to a critical value. Also high IM and IS values should not be considered as they represent a more inaccurate and imprecise fit (Lee *et al.*, 1999).

Analysing IM and IS values from the alpine lakes data-set, reported in Figure 4.5, we can observe a rapid decrease of IM from 3 to 4 number of factors and a further decrease from 5 to 6, with a step between 4 and 5. Instead, IS a first step between 3 and 4 number of factors.

From this analysis it seems that solutions with 3 to 5 number of factors have a better fit.

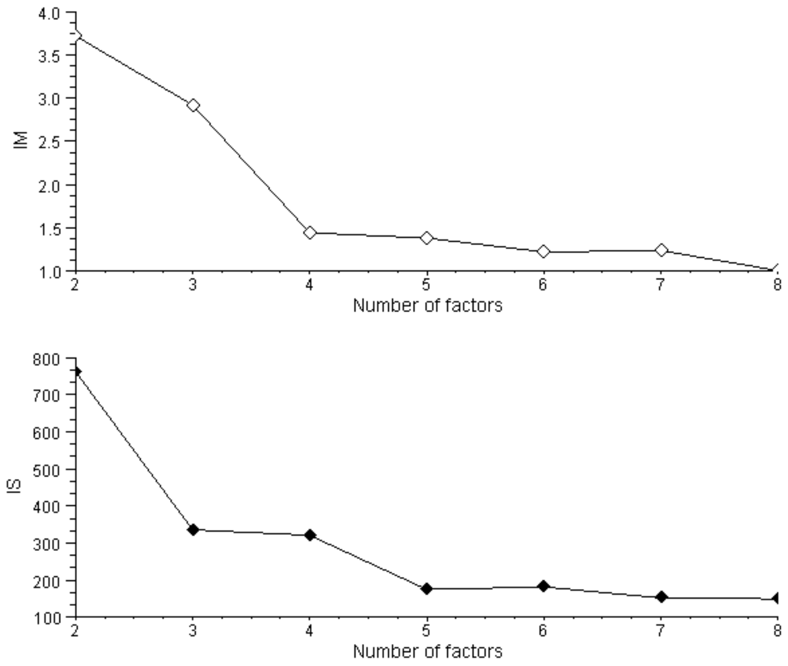


Figure 4.5 – IM and IS plot vs number of factors

4.2.4 Rotmat

The rotmat matrix, indicating the rotational freedom of the solution, give us another tool to elicit information on the number of factors by plotting the maximum element in the matrix (the worst case, corresponding to greater rotational freedom, is used) against the number of factors.

Then we can reject those number of factors from which the maximum element value shows a rapid increase, as they have a high rotational freedom (Lee *et al.*, 1999).

In Figure 4.6, maximum elements of rotmat matrix from the alpine lakes example show a rapid increase for number of factors 2 and 8; these solutions have hence a high rotational ambiguity and, in addition to preceding results, they can be rejected from further analyses.

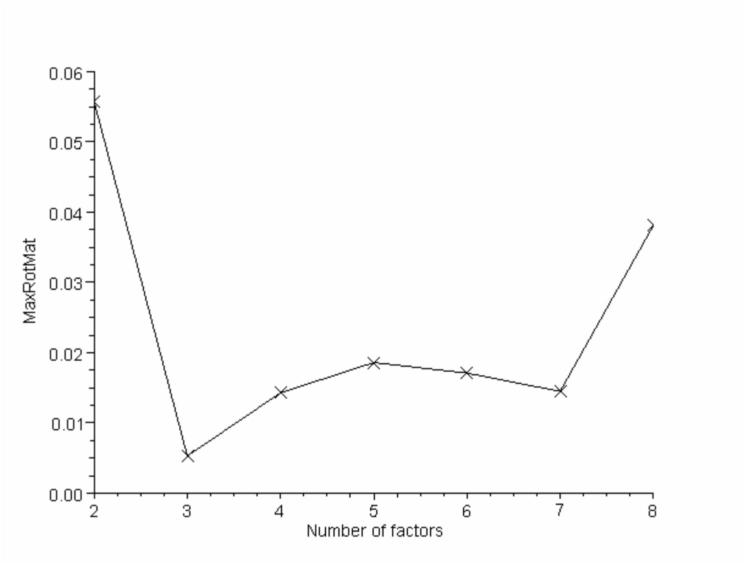
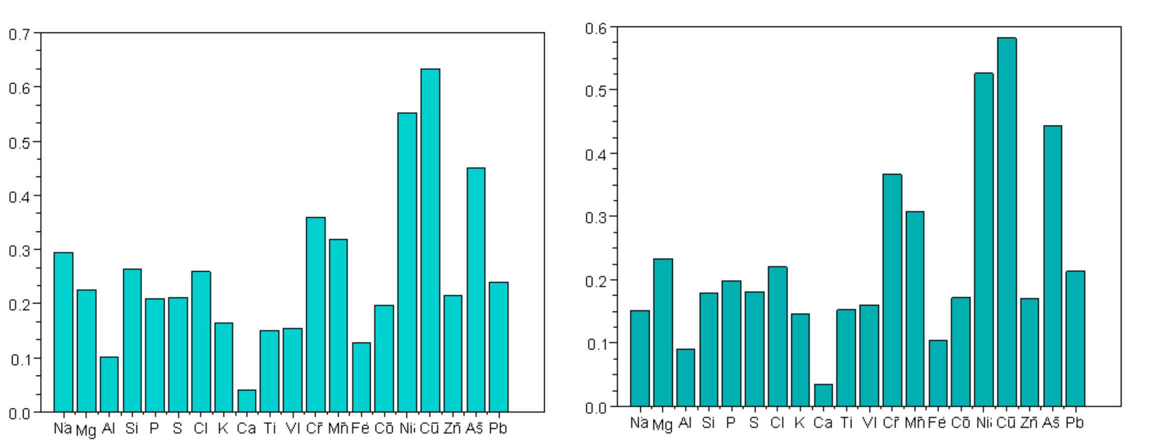


Figure 4.6 – MaxRotMat values vs the number of factors

4.2.5 NEVF

As stated in Section 3.2.4 when a given variable shows NEVF values greater than 0.25 (more than 25% of its variability is not explained by the fit), then the variable is considered not explained by the fit. In this case it is necessary to introduce a further or more factors in order to explain the variable.

NEVF of data from alpine lakes example are reported in Figure 4.7 for solutions with 3, 4 and 5 factors. Moving from 3 to 4 factors, Na and Si NEVF show a greater decrease, reaching values below 25% while other variables record a lower decrease. The 5 factors plot differs from the 4 factor mainly for the variable Cr that decreases its NEVF from 37% to 22%. This great change in Cr NEVF is explained by the isolation of Cr into the new additional factor.



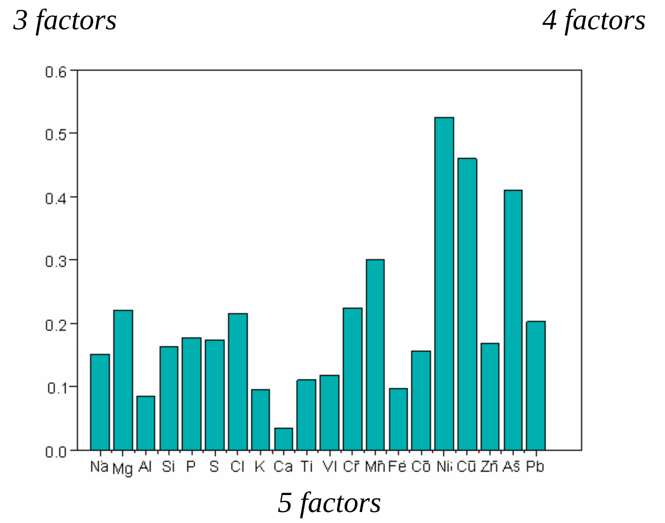


Figure 4.7 – Not Explained Variation of F (NEVF) plots for different number of factor

The variables Ni, As and Cu maintain their NEVF at high values: the reason could be attributed to high percentage of below-detection-limit data, respectively of 59%, 56% and 28%.

4.3 *Controlling rotations*

Usually in the PMF2 algorithm, pseudorandom numbers are used as initial element matrices values. However, when many trials with different rotations have to be performed on the same dataset in order to evaluate all the possible solutions, the use of pseudorandom number do not seems to be the right choice. This is because different local minima may be produces with this type of initialization respect to the selected rotation and also the calculated factors may appear with a different index in each solution, making the comparison among solutions more complicated. Since these effects might mask the rotational effects on the solutions (Paatero *et al.*, 2002), Paatero suggests the following scheme when operating with rotations:

- Perform different initialization runs with pseudorandom value and $\phi = 0$ (central solution) in order to evaluate the Q stability.
- Choose the best central solution and use it as a starting point for the data processing with rotations. This is done using the "*goodstart*" parameter as described in Appendix (Rotations).

- For a complete point of view this step may be repeated with another central solution as a starting point.

With the above procedure the effects produced by the rotations on the solutions are clearest to compare.

Once different runs have been made and different numbers of factors and rotations have been explored, it needs to reject the solutions that do not satisfy some criteria. One of these evaluation techniques is the Q-value investigation respect the FPEAK parameter, ϕ.

Before describing this method an apparent contradiction has to be remarked: when analyzing the Q versus ϕ dependence it may happen that the solution with a non-zero ϕ shows a slightly higher Q. This might result in contradiction with the rotation definition in make unchanged the factor matrices product ($GF = GTT^{-1}F$). Actually the G and F factors are "*flexible*" that is their product can differ a little bit from the rotate factor product ($GF \approx GTT^{-1}F$) and so a worse fit has been accepted in order to minimize the object function. This is due by the non-negative forcing of the matrices elements and it is said that a distorted rotation is performed (Paatero *et al.*, 2002).

In conclusion, also solutions with Q value that is not too high than the central one ($\phi = 0$) have to be considered.

Based on years of experience, customary trend of Q respect to the ϕ value is described in the article by Paatero *et al.*, 2002. Starting from the central solution and observing the behaviour of Q related to the increase of the ϕ parameter, two distinct phases may be distinguished. In the first phase the Q value grow slowly while in the latter one the Q value increases very quickly and the factor matrices tend to be distorted because of the non-negativity constraint. In seems that useful results appear when ϕ is near the end of the first phase; however further experience is needed in order to have a best knowledge in choosing ϕ values. Anyway, this could be a helpful tool to make a first step decision on the rotate solutions to be considered.

4.3.1 Assessing the increase of Q

As previously mentioned, rotations are considered also if the Q value is not too different from the central solution one. In order to quantify the expression “not too different” we can compare the Q value computed from the fit with the Q^{exp} or the Q of the central solution.

When a rotation is performed and some of the G or/and F matrix elements move to near zero value (say *z* the number of these elements), then Q^{exp} value increase because of the near zero elements are viewed as non free parameter. The Q^{exp} increment is equal to *z* number. Clearly it is not possible to define a precise rule, based on Q value, that allow us to decide when a rotation is to rejected, but, as a practical decisional step we could considered forbidden rotations that show an increase of Q values, respect to the central solution one, above than 10% (Paatero *et al.*, 2002).

In Figure 4.8 an example of the variations of Q values respect the FPEAK parameter. The plot is referred to the 5 factors solution of the alpine lakes data-set and reports the ratio between the calculated Q value and the Q value from the central solution (Qzero). All the ratios are closed to the unit but moving toward too high positive and negative FPEAK the ratio increases.

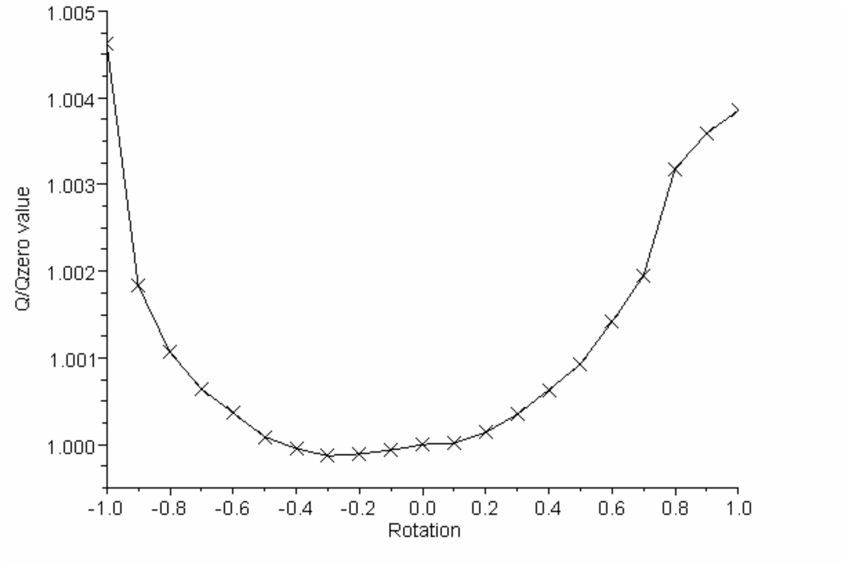


Figure 4.8 – Q/Qzero values from different rotations (Qzero is the Q value corresponding to the central solution)

4.3.2 Scaled residual

As in the case of number of factors, the scaled residual can be used to reject some rotations. Observing the residual plot it is possible to detect the rotations that show a better residual distribution. However, as already explained some deviation from the expected distribution may be due to not well known standard deviation, so it is advisable to control them.

4.3.3 IM, IS and rotmat

The parameters IM, IS and rotmat, described in the previous section, are used to select the most meaningful range from all the FPEAK values input into the model. The range of interest should have low and stable IM and IS values, representing the more accurate and precise fits.

In Figure 4.9, from the 5 factors solution of alpine lakes data-set, IM shows lower values in the range -0.4 – 0.2, while for IS a continuous decrease until 0.7 take place.

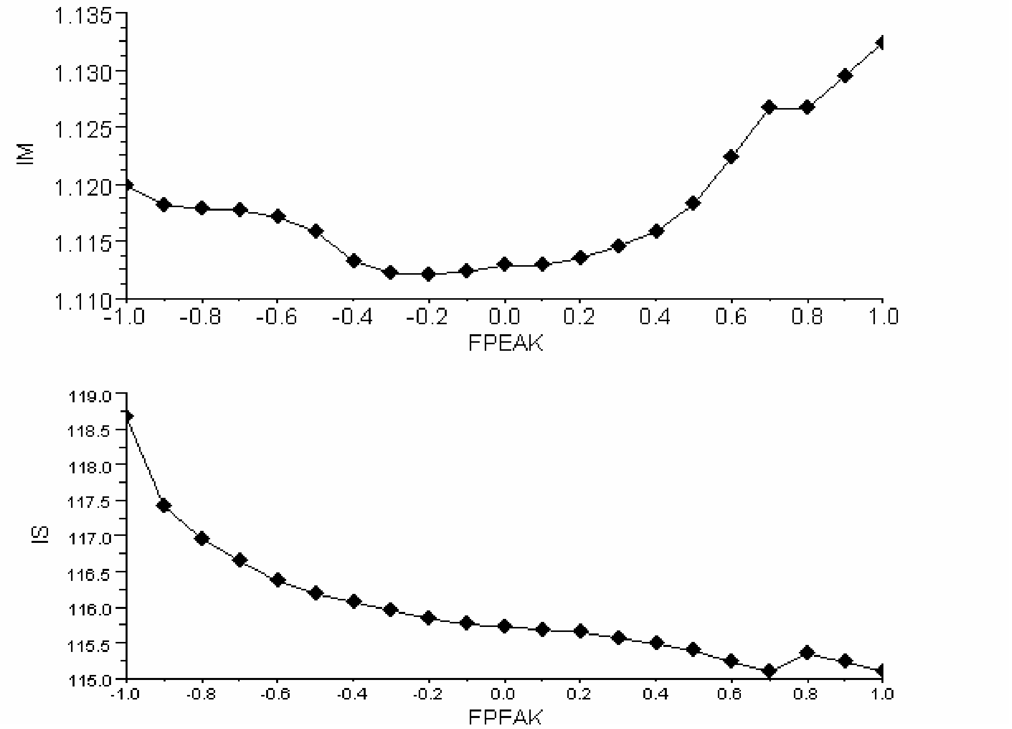


Figure 4.9 – IM and IS valued for different FPEAK

Using the rotmat matrix, we find the FPEAK values that show lower maximum element of rotmat, corresponding to a lower rotational freedom; therefore these rotations will be favored (Lee *et al.*, 1999).

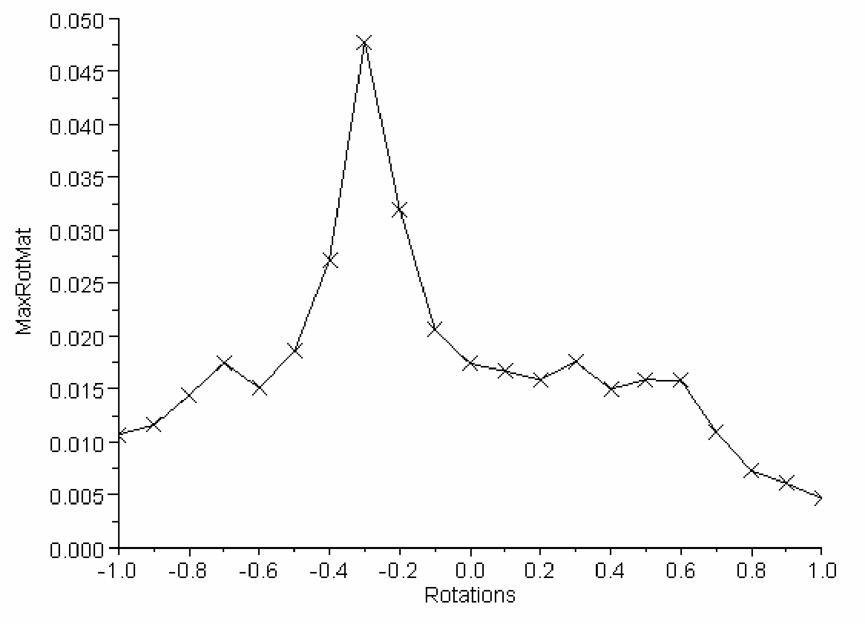


Figure 4.10 – RotMat values from different FPEAK

Rotations with FPEAK from -0.5 to to 0.1 show higher rotational ambiguity (Figure 4.10) and can be rejected while, in addiction to IM and IS results, FPEAK values from 0 to 0.2 are favoured.

4.3.4 Fkey: a *priori* information

An alternative for controlling rotations is the use of a priori information (Paatero *et al.*, 2002). Selection among different solutions given by different ϕ (FPEAK) values may be performed by the knowledge of some information on the problem under analysis, extracted from preceding studies; this allows users to reject non representative rotations. A priori information may be input within the algorithm through the use of Fkey matrix (see Appendix, Rotations) that pulls down to zero some F elements. Like this, Fkey matrix guides the analysis towards a more understanding solution/rotation. In example, if it is known a priori that one or more variables have a zero contribution on some factors, this information can be implemented in order to force the variable to the known values (see a Fkey example applied to the analysis of atmospheric particulate matter on

Lee *et al.*, 1999). However forcing to zero elements in F matrix seems to increase the frequency of local minima, giving rise to multiple problem solution (Paatero, 1997); this problem may however be overcome processing different run, starting from various pseudorandom values.

An interesting application on the use of a priori information in the contest of atmospheric pollution is describe in Lingwall and Christensen, 2007. They studied the priori information effects, with Fkey and the target source profile (Gkey, see Appendix, Rotations), using simulated experiments on ambient air pollution data and varying the correctness of such information. In general they found out an improved source profiles and source contribution when the pulling to zero elements is performed on 'clean data' (i.e. data with low uncertainties and not affected by unidentified source), otherwise the results are nearly the same. However the fit was worsened if the information carried by the zero elements was not correct. In the case of the target source profile method, the estimate of source profile and contribution was improved even if the profiles have some inaccuracy.

Therefore it is advisable to do not introduce many a priori information into the algorithm especially if we are not sure of their trueness. This may lead in fact to a poorer analysis of the problem because of a priori rejecting of some solutions/rotations.

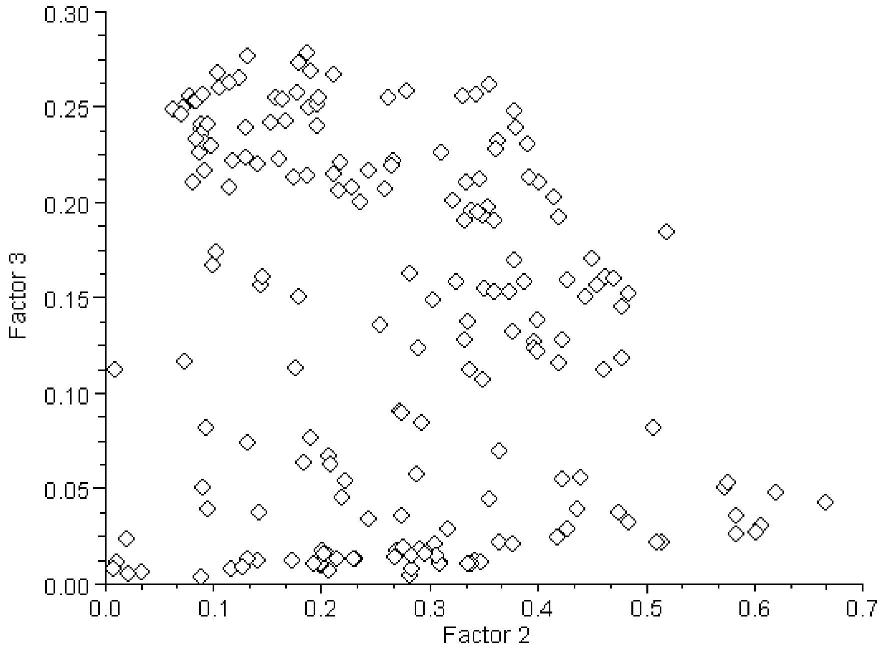
We suggest, firstly, running the model without the implementation of a priori information and, after solutions are obtained, using such information to reject ones of them. Subsequently, solutions may be recalculated, adding a priori information within the algorithm in order to compare the results of these two different methods.

4.3.5 A graphical diagnostic method

The graphical approach is a simple method to assessing the optimal rotation, also called *G space plotting for PMF modelling* (Paatero *et al.*, 2005). This procedure is carried out after the model is run and a number of factors is chosen. It is made the assumption that the determined factors are uncorrelated each other. Actually there is always a weak correlation between pairs of factors, called *weak independence*. The goal of this method is to reject the rotation that give correlation between pair of factor and to show this, G_{ik} elements corresponding to two different factors are plotted in a Cartesian plane. All the points lie in the positive quadrant because of the non-negative constraint and, if the factors are not correlated, the straight lines passing thought the origin of axes and including all the point between them should approximate the Cartesian axes. These lines, called *edges*, split the positive quadrant in two regions: one contains all the points and the other with no points (or with outliers that lies away). Hence, the factor plotting that have the edges nearest the axes are those relating to the optimum rotation.

However it is important to note that there may be physical situations where oblique edges can occur and so the optimal rotation is not identify with this method; priori information, if available, can help interpreting the plot. Also edges parallels to Cartesian axes do not guarantee that the solution is unique! (Paatero *et al.*, 2005).

In Figure 4.11 an example of two G plots from the alpine lakes analysis is reported. In the first one it is clear that factor 2 and factor 3 are uncorrelated each other, while the second show some correlation between factor 4 and factor 5. The correlation may indicate either an uncorrected rotation or a natural trend, and a good knowledge of the problem could be helpful to make a decision.



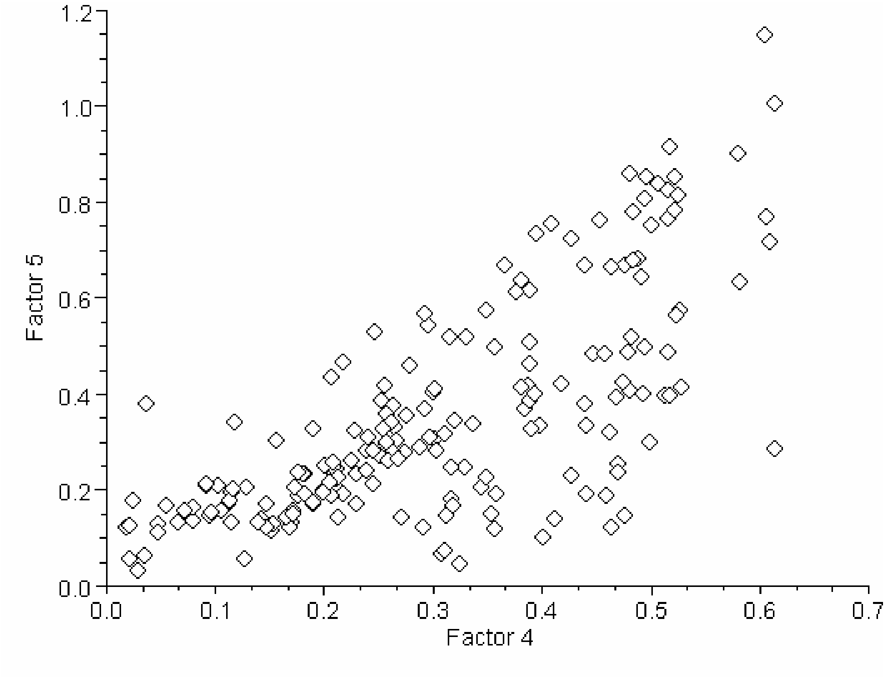


Figure 4.11 – G plot between factor 2 and 3 (up) and factor 4 and 5 (below)

4.4 *Non-linear variables*

As already explained, PMF is a linear model respect to the factors and generally the input variables obey this property, i.e. chemical elements concentrations are additive when emitted by different sources. However certain variables, such as pH, are non-linear (as in the example, pH values are not additive when emitted by different sources). In the article by Reinikainen, 2001, relate to the study on water quality in Lake Saimaa, this problem is treated by using the expression 7.5-pH instead the variable pH. The new variable is even now non linear but it has the property to increase when the acidifying emission increase. For a more careful factors estimate the pH uncertainties were increased.

Appendix

Guide for PMF programs

Paatero PMF2

PMF2 program runs under DOS environment (it is not an installation program).It uses an initialization file with extension .INI in order to input desired parameters into the algorithm. In this file the users can select the input file (data and uncertainties matrices) and all the parameters and information that will be used in the problem under analysis.

.INI file

In this section the general run control parameters, which can be set in the .INI file to determine the optimum solution, are described. See Paatero, 2004a and Paatero, 2004b user's guide for a more detailed discussion on the .INI file.

Monitor:

This parameter controls the number of monitoring output produced by the algorithm. With the default value, set to 1, every step is reported on screen and in .log file. If monitor>1 then only the Mth step is reported.

Dimensions:

In this part of the code the number of rows and columns of the data and uncertainties matrices (they must have the same dimensions) and the number of factors must be inserted. For a more quickly computation it should be better to have rows number > columns number.

Repeats:

Set the repeat value equal to the number of different repeated computation to do in each run. Between different repeats some information may be varied, in examples starting from different initialization numbers. The information to be recalculated at each run must have the (R) code in the Input/Output table set to T or true value. The .INI file is read only once.

Fpeak:

A positive or negative value implements rotation. The central solution is achieved with FPEAK=0 (default value).

Robust mode:

When set to true value (default value) the algorithm take into account the outliers data using the enhanced Q described in Section 3.2.1 and the here below listed outlier distance.

Outlier distance

This parameter is set in order to select a threshold to identify outliers. The following values are suggested: α = 0.2, 0.4 (default value), 0.8. Also the program enables to set two different thresholds, one for positive residues, α_p, and the other for negative residues, α_n, by the use of the optional parameter *outlimits* α_p α_n.

Codes C1, C2, C3 and Errormodel

The three codes and the Errormodel (EM) are used as explained in Section 3.2.2 in order to determine the data standard deviation when the uncertainties matrix is not input.

Pseudorandom Seed

According to the seed value, pseudorandom numbers are generated to initialize the algorithm.

Iteration control table

The model convergence can be controlled by means of four parameters, each of them having three values corresponding to three subsequent convergence steps. The third level is more involved in the fit convergence as the first two steps are only used to address the initial values of the factor matrices toward a more realistic solution; so it is important to make the right choice for this last level. Below, the explanation of the parameters and the common used value:

- *Lims*: it represents a weight coefficient for the penalty and regularization terms and its value indicates the closeness of the matrices component to zero values. As to the last step, the value can be chose in relation to the number of data point. For large model a good range is 1.0 – 10.0, while for small model can be used a lower value as 0.01;
- *Chi2_test*: represent a threshold fixed for the fit convergence;

- *Ministep_required*: number of convergence consecutive steps (say N);
- *Max_cumul_count*: maximum number of cumulative steps allowed for the convergence.

Once fixed these values, the fit is said to converge if the variability of Q value (χ^2) is lower than *Chi2_test* after N consecutive step without go over the maximum number of cumulative steps allowed (*Max_comul_count*).

It is interesting to test if different parameter values produce similar results; like this we are sure of a good convergence.

Rotations

In PMF2 rotations can be induced by four different techniques (Paatero, 2004a).

As to the first, rotations can be implemented in the model by the FPEAK value (ϕ) with a zero value default setting (central solution). This method is simple because the users have only to select the desired rotational value. In order to examine different rotations it is better to start from lowest ϕ values and use, as a starting point for the following computation, the result obtained from the previous computation. This is done using the optional parameter *goodstart* (do not use the *sortfactors* parameters, see following subsection).

The second method to induce rotation uses the matrix *rotocom*, but this technique is not recommended and not yet use in practical applications.

The third method is based on a priori knowledge of information about the problem under analysis when some elements in the factor matrix F are known to have zero or very small values. Pulling down elements is done acting on the penalty term of the Q function by means the *Fkey* matrix, an integer values matrix with the same dimensions of F and which controls the corresponding F elements behavior. The Fkey matrix contains all zero values except in correspondence to the pulling-down elements; for these points values are as greater as the elements are likely to be zero. The influence of Fkey is exponential (Lee *et al.*, 1999). It is obvious that, as to the first, a Fkey = 0 solution must be found in order to detect if the a priori information are already satisfied by the fit.

To enable Fkey matrix, the FIL code (= 0 is the default value) in the .INI file must be set to a non zero value; it is recommended to use FIL = 4 to include this matrix in the .INI file.

The latter method uses target factor shape and it is the most complicated method. At the time no experiences have been found about its usefulness. This approach is based on “pseudo measurements” that are included in the X matrix and which represents the target shapes for each factor. New rows are added in the X and standard deviation matrices in a number equal to the number of factors and the priori information about the factor shapes are used to set them. A *Gkey* matrix is used like the Fkey.

Robuste method

The so called robust mode may be activated or not and it is also possible to select the threshold distance used to handle outliers (the default value is set to 4.0).

Multiple results

Computing different initial runs, using the parameter “*Repeats*”, is necessary to assess the Q stability. In fact the Q expression may have one or more local minima, and it may even happen that the optimum solution does not correspond to the global one. The correct statistical handling of this situation is not known yet, but there are indications that local minima tend to occur when too few or too many factors are used in the algorithm.

In order to detect if there are not local minima several runs must be performed, starting from different point. This can be done either changing to each run the *SEED* value, used to set different pseudorandom starting point, or in the simplest way, raising the *repeat* number from the default value (=1) and in addition activating the repeated runs putting the ‘true’ (T) value in the (R) codes of the factor matrices.

Customary, results from different runs show the factors in random order and this make difficult to compare the obtained results. Using the optional command *sortfactorsg* or *sortfactorsf*, the factors appear about in the same position from run to run.

Normalization of factor matrices

If necessary, near the bottom of the .INI file, the user can normalize the factor matrices according to the following options (Paatero, 2004a):

- None: no normalization
- MaxG = 1: the maximum absolute value in each G column is equal to the unity
- Sum|G| = 1: the sum of elements absolute value in each G column is equal to the unity
- Mean|G| = 1: the mean value of elements absolute value in each G column is equal to the unity
- MaxF = 1: the maximum absolute value in each F row is equal to the unity
- Sum|F| = 1: the sum of elements absolute value in each F row is equal to the unity
- Mean|F| = 1: the mean value of elements absolute value in each F row is equal to the unity

With one of these operations the GF product does not change because of columns of G and rows of F are respectively divided and multiplied by the same p normalization coefficients.

Output files

The outputs produced by the program and organize into .txt file according to the users preferences are: G and F factor matrices, G and F standard deviations, copy of the input matrices, scaled residual matrix, explained variation (EV) matrix of G and F, Q value of the fit and the *rotmat* matrix. A .LOG file is also produced in which possible errors made by the algorithm are reported

The produced residuals (i.e. the difference between the measured ad the fitted values) are useful tools to gain information on the Q values and the quality of the standards deviation if the latter are not well known (see Section 4.2.2)

Rotomat is a *pxp* matrix describing the degrees of rotation of the results that is a measure of the rotational ambiguity of the obtained solution. Applying FPEAK to the central solution there will be less rotational ambiguity for further rotations; correspondingly, rotomat values are decreasing. Rotomat values must only be purely indicative (Paatero, 2004a)

With regard to the G and F standard deviation matrices, their values are not too accurate if the uncertainties of X elements are not correct as they are computed starting from X matrix and the uncertainties matrix.

With regard to the G and F standard deviation matrices, their values are not too accurate if the input uncertainties are not well-known and hence computed with one of the method described in the Chapter 3.

Bibliography

Anttila P., Paatero P., Tapper U., Järvinen O., (1995). Source identification of bulk wet deposition in Finland by positive Matrix Factorization. *Atmospheric Environment*, **14**, 1705–1718.

Begum B.A., Kim E., Biswas S.K., Hopke P.K., (2004). Investigation of sources of atmospheric aerosol at urban and semi-urban areas in Bangladesh. *Atmospheric Environment*, **38**, 3025-3038.

Bzdusek P.A., Christensen E.R., Lee C.M., Pakadeesusuk U., Freedman D.C., (2006). PCB congeners and dechlorination in sediments of Lake Hartwell, South Carolina, determined from cores collected in 1987 and 1988. *Environmental Science and Technology*, **40**, 109–119.

Lee E., Chan C.K., Paatero P., (1999). Application of Positive Matrix Factorization in source apportionment of particulate pollutants in Hong Kong. *Atmospheric Environments*, **33**, 3201-3212.

Chang C-C., Wang J-L., Lung S-C., Liu S-C., Shiu C-J., (2009). Source characterization of ozone precursors by complementary approaches of vehicular indicator and principal component analysis. *Atmospheric Environment*, **43**, 1771–1778.

Critto A., Carlon C., Marcomini A., (2003). Characterization of contaminated soil and groundwater surrounding an illegal landfill (S. Giuliano, Venice, Italy) by principal component analysis and kriging. *Environmental Pollution*, **122**, 235–244.

DelValls T.A., Forja J.M., González-Mazo E., Gómez-Parra A., (1998). Determining contamination sources in marine sediments using multivariate analysis. *Trends in analytical chemistry*,**17**, 181–192.

Dos Santos J. S., De Oliveira E., Bruns R. E., Gennari R.F., (2004). Evaluation of the salt accumulation process during inundation in water resource of Contas river basin (Bahia–Brazil) applying principal component analysis. *Water Research*, **38**, 1579–1585.

Ellison S.L.R., Rosslein M., Williams A. (2000). Quantifying Uncertainty in Analytical Measurement (QUAM) - Second Edition. *EURACHEM/CITAC.*

Free G., Solimini A.G., Rossaro B., Marziali L., Giacchini R., Paracchini B., Ghiani M., Vaccaro S., Gawlik B., Fresner R., Santner G., Schönhuber M., Cardoso A.C.. Modelling the response of lakes macroinvertebrate species in the shallow sub-littoral: Relative roles of habitat, lake morphology, aquatic chemistry and sediment composition.

Hopke P.K. (a). Receptor models for particulate matter management. *Department of Chemistry, Clarkson University, Potsdam, NY 13699-5810*.

Hopke, P.K., 2000. A guide to positive matrix factorization. EPA Workshop Proceedings Materials from the Workshop on UNMIX and PMF as Applied to PM2.5.

Hopke P.K., (2003). Recent developments in receptor modeling. *Journal of chemometrics*, **17**, 255–265.

Huang S., Rahn K.A., Arimoto R,. (1999). Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island. *Atmospheric Environment*, **33**, 2169–2185.

ISO/IEC, (2008). Guide to the expression of uncertainty in measurement (GUM)

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

Kim E., Hopke P.K., (2007). Comparison between sample-species specific uncertainties and estimated uncertainties for the source apportionment of the speciation trends network data. *Atmospheric Environment*, **41**, 567–575

Lingwall J., Christensen W.F., (2007). Pollution source apportionment using a *priori* information and positive matrix factorization. *Chemometrics and Intelligent Laboratory Systems*, **87**, 281–294.

Lee E, Chan C.K., Paatero P., (1999). Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. *Atmospheric Environment*, **33**, 3201–3212.

Loska K., Wiechuła D., (2003). Application of principal component analysis for the estimation of source of heavy metal contamination in surface sediments from the Rybnik Reservoir. *Chemosphere*, **51**, 723–733.

Lu J., Jiang P., Wu L., Chang A.C., (2008). Assessing soil quality data by positive matrix factorization. *Geoderma*, **145**, 259-266.

Motelay-Massei A., Ollivon D., Garban B., Chevreuil M., (2003). Polycyclic aromatic hydrocarbons in bulk deposition at a suburban site: assessment by principal component analysis of the influence of meteorological parameters. *Atmospheric Environment*, **37**, 3135–3146.

Norris G., Vedantham R., Wade K., Brown S., Prouty J., Foley C., (2008). EPA Positive Matrix Factorization (PMF) 3.0: fundamentals & user guide. *U.S. Environmental Protection Agency*

Officer S.J., Kravchenko A., Bollero G.A., Sudduth K.A., Kitchen N.R., Wiebold W.J., Palm H.L., Bullock D.G., (2004). Relationships between soil bulk electrical conductivity and the principal component analysis of topography and soil fertility values. *Plant and Soil*, **258**, 269–280.

Paatero P. (a). Noisy variables in factor analytic models. *University of Helsinki, Dept. Physical Sciences, Finland*

Paatero P., (1997). Least square formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, **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. *Journal of Graphical Statistics*, **8**, 854–888.

Paatero P., (2004a). User’s guide for positive matrix factorization programs PMF2 and PMF3, Part1: tutorial. *University of Helsinki, Helsinki, Finland.*

Paatero P., (2004b). User’s guide for positive matrix factorization programs PMF2 and PMF3, Part2: references. *University of Helsinki, Helsinki, Finland*.

Paatero P., (2004). Introduction to PMF – positively constrained factor analysis with individual weighting of matrix elements. *ftp://ftp.clarkson.edu/pub/hopkepk/pmf/*.

Paatero P., (2007). End User’s Guide to Multilinear Engine Applications. *ftp://ftp.clarkson.edu/pub/hopkepk/pmf/*.

Paatero P., Hopke P.K., Song X.-H., Ramadan Z., (2002). Understanding and controlling rotations in factor analytic models. *Chemometrics and Intelligent Laboratory Systems*, **60**, 253–264.

Paatero P., Hopke P.K., (2003). Discarding or downweighting high-noise variables in factor analytic models. *Analytica Chimica Acta*, **490**, 277–289.

Paatero P., Hopke P.K., Begum B.A., Biswas S.K., (2005). A graphical diagnostic method for assessing the rotation in factor analytical models of atmospheric pollution. *Atmospheric Environment*, **39**, 193–201.

Paatero P., Tapper U,. (1993). Analysis of different modes of factor analysis as least squares fit problems. *Chemometrics and Intelligent Laboratory System*, **18**, 183–194.

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

Peré-Trapat E., Ginebreda A., Tauler R., (2007). Comparison of different multiway methods for the analysis of geographical metal distributions in fish, sediments and river waters in Catalonia. *Chemometrics and Intelligent Laboratory Systems*, **88**, 69–83.

Pires J.C.M., Sousa S.I.V., Pereira M.C., Alvim-Ferraz M.C.M., Martins F.G., (2008). Management of air quality monitoring using principal component and cluster analysis-Part I: SO_2 and PM_{10}. *Atmospheric Environment*, **42**, 1249–1260.

Polissar A.V., Hopke P.K., Malm W.C., Sisler J.F., (1998). Atmospheric aerosol over Alaska: 2. Elemental composition and sources. *Journal of Geophysical Research,* **103**, 19 045–19 057.

Polissar A.V., Hopke P.K., Paatero P., Kaufmann Y.J., Hall D.K., Bodhaine B.A., Dutton E.G., Harris J.M., (1999). The aerosol at Barrow, Alaska: long-term trends and source locations. *Atmospheric Environment*, **33**, 2441–2458.

Polissar A.V., Hopke P.K., Poirot R.L., (2001). Atmospheric aerosol over Vermont: chemical composition and sources. *Environmental Science and Technology*. **35**, 4604–4621.

Reinikainen S.-P., Laine P., Minkkinen P., Paatero P., (2001). Factor analytical study on water quality in Lake Saimaa, Finland. *Journal of Analytical Chemistry*. **369**, 727–732.

Reimann C., Filzmoser P., Garrett R.G., (2002). Factor anlysis applied to regional geochemical data: problems and possibilities. *Applied Geochemistry*, **17**, 185–206.

Sakihama H., Ishiki M., Tokuyama A., (2008). Chemical characteristics of precipitation in Okinawa Island, Japan. *Atmospheric Environment*, **42**, 2320–2335.

Singh K.P., Malik A., Singh V.K., Sinha S., (2006). Multi-way data analysis of soils irrigated with wastewater. A case study. *Chemometrics and Intelligent Laboratory Systems*, **83**, 1-12.

Song X.H., Polissar A.V., Hopke P.K., (2001). Sources of fine particle composition in the northeastern US. *Atmospheric Environment*, **35**, 5277–5286.

Soonthornnonda P., Christensen E.R., (2008). Source apportionment of pollutants and flows of combined sewer wastewater. *Water Research*, **42**, 1989–1998.

Unonius L., Paatero P., (1990). Use of Singular Value Decomposition for analysing repetitive measurements. *Computer Physics Communications*, **59**, 225–243.

Vaccaro S., Sobiecka E., Contini S., Locoro G., Free G., Gawlik B.M., (2007). The application of positive matrix factorization in the analysis, characterization and detection of contaminated soils. *Chemosphere*, **69**, 1055-1063.

Watson J.G., Chen L.W.A., Chow J.C., Doraiswamy P. Lowenthal D.H., (2008). Source Apportionment: findings from the U.S. Supersite Program. *Technical paper, Air&Waste Manage. Assoc.*, **58**, 265–288.

Xie Y., Berkowitz C.M., (2006). The use of positive matrix factorization with conditional probability functions in air quality studies: an application to hydrocarbon emissions in Houston, Texas. *Atmospheric Environment*, **40**, 3070–3091.

Xie Y., Hopke P.K., Paatero P., Barrie L.A., Li S.-M., (1998). Identification of source nature and seasonal variations of Arctic aerosol by the multilinear engine. *Atmospheric Environment*, **33**, 2549–2562.

Yu T.-Y., Chang L.-F. W., (2000). Selection of the scenarios of ozone pollution at southern Taiwan area utilizing principal component analysis. *Atmospheric Environment*, **34**, 4499–4509.

List of Abbreviations

BDL: Below Detection Limit

CMB: Chemical Mass Balance

EM: Error Model

EVF: Explained Variation of F

EVG: Explained Variation of G

FA: Factor Analysis

LS: Least Square

ME: Multilinear Engine

MV: Missing Value

NEVF: Not Explained Variation of F

NEVG: Not Explained Variation of G

PCA: Principal Component Analysis

PM: Particulate Matter

PMF: Positive Matrix Factorization

SVD: Single Value Decomposition

TTFA: Target Transformation Factor Analysis

XRF: X Ray Fluorescence

European Commission

EUR 23946 EN – Joint Research Centre – Institute for Environment and Sustainability
Title: Positive Matrix Factorisation (PMF) - An Introduction to the Chemometric Evaluation of Environmental Monitoring Data Using PMF
Author(s): Sara Comero, Luisa Capitani, Bernd Manfred Gawlik
Luxembourg: Office for Official Publications of the European Communities
2009 – 58 pp. – 21.0 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-12954-4
DOI 10.2788/2497

Abstract

Positive Matrix Factorization (PMF) is a multivariate factor analysis technique used successfully among others at the US Environmental Protection Agency for the chemometric evaluation and modelling of environmental data sets. Compared to other methods it offers some advantage that consent to better resolve the problem under analysis. In this report, the algorithm to solve PMF and the respective computer application, PMF2, is illustrated and, in particular, different parameters involved in the computation are examined.
Finally, a first application study on PMF2 parameters setting is conducted with the help of a real environmental data-set produced in the laboratories of the JRC Rural, Water and Ecosystem Resource Unit.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LB- NA- 23946- EN- C

Publications Office
Publications.europa.eu

ISBN 978-92-79-12954-4

9 789279 129544