ELSEVIER

Chemometrics and Intelligent Laboratory Systems 60 (2002) 253–264

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Understanding and controlling rotations in factor analytic models

Pentti Paatero[a], Philip K. Hopke[b,*], Xin-Hua Song[b], Ziad Ramadan[b]

[a]*Department of Physics, University of Helsinki, PO Box 64, FIN-00014 Helsinki, Finland*
[b]*Department of Chemical Engineering, Clarkson University, Box 5705, Potsdam, NY 13699-5705, USA*

Abstract

Positive Matrix Factorization (PMF) is a least-squares approach for solving the factor analysis problem. It has been implemented in several forms. Initially, a program called PMF2 was used. Subsequently, a new, more flexible modeling tool, the Multilinear Engine, was developed. These programs can utilize different approaches to handle the problem of rotational indeterminacy. Although both utilize non-negativity constraints to reduce rotational freedom, such constraints are generally insufficient to wholly eliminate the rotational problem. Additional approaches to control rotations are discussed in this paper: (1) global imposition of additions among "scores" and subtractions among the corresponding "loadings" (or vice versa), (2) constraining individual factor elements, either scores and/or loadings, toward zero values, (3) prescribing values for ratios of certain key factor elements, or (4) specifying certain columns of the loadings matrix as known fixed values. It is emphasized that application of these techniques must be based on some external information about acceptable or desirable shapes of factors. If no such a priori information exists, then the full range of possible rotations can be explored, but there is no basis for choosing one of these rotations as the "best" result. Methods for estimating the rotational ambiguity in any specific result are discussed.

Keywords: Multilinear engine; Factor analysis; Receptor models; Source apportionment

1. Introduction

One of the significant problems in the use of factor analysis to solve the mixture resolution problem is that there is rotational indeterminacy in the solution. Henry [1] presents this issue in terms of the factor analysis problem being "ill-posed". The problem can be solved, but it does not produce a *unique* solution.

To illustrate this problem, Fig. 1 shows simulated data for ambient samples that consist of mixtures of the soil and basalt source profiles [2]. This figure shows a plot of the iron and silicon values for a series of simulated samples. There need to be two profiles to reproduce each data point, but these "profiles" could range from the original axes to any of the other pairs of lines. Because there are no zero valued source contributions, the solid lines are not the "true" profiles. The true source profiles lie somewhere between the inner solid lines that enclose all of the points and the original axes, but without additional information, these profiles cannot be fully determined.

This paper discusses rotations and uniqueness of the bilinear non-negatively constrained factor analytic model. Part of the discussion is independent of the technique used to calculate the factor model, and part applies specifically to Positive Matrix Factorization (PMF) [3]. Many of the conclusions and recommendations are based on the extensive hands-on experi-

[*] Corresponding author. Tel.: +1-315-268-3861; fax: +1-315-268-6654.
E-mail address: hopkepk@clarkson.edu (P.K. Hopke).

0169-7439/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(01)00200-3

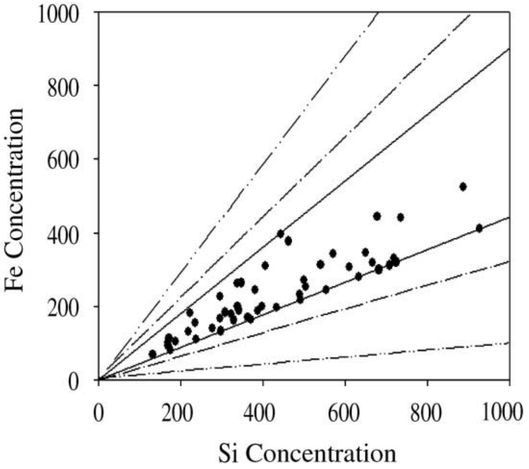

Fig. 1. Simulated data showing multiple possible source "profiles" that could be used to fit the data.

ence of authors when analyzing various data sets rather than exact mathematical or statistical analyses. Statistical analyses are severely hampered because of the non-linear nature of the bilinear model and because the distributions and covariances of errors of data values are not known in practice. In particular, the assumption of statistically independent errors would be a gross over-simplification for most environmental multivariate data sets.

By using the element-wise notation, the bilinear factor analytic model is written as:

$$x_{ij} = \sum_{h=1}^{p} g_{ih} f_{hj} + e_{ij}$$
$$(i = 1, \ldots, m, \quad j = 1, \ldots, n). \tag{1}$$

Customarily, the elements g_{ih} of matrix **G** are called scores and elements $\boldsymbol{f}_{hj}$ of **F** are called loadings. However, the model is fully symmetric with respect to an interchange of **G** and **F**. Hence, the notation of scores and loadings is arbitrary; it should just be taken as a convenience. In atmospheric studies, i is time index while j is variable index. In spectroscopy, i would be the wavenumbers index while j enumerates samples.

In matrix notation, the same system of equations is written as:

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

In many environmental and chemometric applications, it is required that all factor elements be non-negative: $g_{ih} \geq 0, f_{hj} \geq 0$. Throughout this work, it is assumed that these constraints are imposed, unless noted otherwise.

The "loss function" Q is defined as follows:

$$Q = Q(\mathbf{G}, \mathbf{F}) = \sum_{i=1}^{m} \sum_{j=1}^{n} (e_{ij}/s_{ij})^2 \tag{3}$$

where the values s_{ij} are error estimates of data values. The unknown "factor elements" g_{ih} and f_{hj} are to be determined so that the loss function Q is minimized.

The number of factors, p, is assumed known throughout this paper. In practice, deciding upon the best value for p is often difficult. This question has been discussed in literature [4,5] and is hence omitted from the present work. In practice, one usually has to repeat the analysis with different values of p. One chooses the solution that appears most useful.

1.1. Rotations and uniqueness

The pair of factor matrices (**G**,**F**) can be transformed to another pair ($\bar{\mathbf{G}}$,$\bar{\mathbf{F}}$) so that the same Q corresponds to the pairs (**G**,**F**) and ($\bar{\mathbf{G}}$,$\bar{\mathbf{F}}$). Then it is said that the first pair can be rotated so that it becomes equal to the second pair. Assume that **T** is a nonsingular square matrix of dimensions $p \times p$. Then **T** has an inverse matrix $\mathbf{T}^{-1}$ so that $\mathbf{T}\ \mathbf{T}^{-1} = \mathbf{I}$. A transformation is defined with the help of **T** by setting:

$$\bar{\mathbf{G}} = \mathbf{GT} \text{ and } \bar{\mathbf{F}} = \mathbf{T}^{-1}\mathbf{F}. \tag{4}$$

Then the products of the pairs (**G**,**F**) and ($\bar{\mathbf{G}}$,$\bar{\mathbf{F}}$) are equal because $\bar{\mathbf{G}}\bar{\mathbf{F}} = \mathbf{GTT}^{-1}\mathbf{F} = \mathbf{GIF} = \mathbf{GF}$. This shows that the transformation defined by Eq. (4) is a rotation. However, this rotation is only acceptable or allowed if all elements of the new factor matrices are also non-negative: $\bar{g}_{ih} \geq 0, \bar{f}_{hj} \geq 0$. The non-negativity constraints limit the rotations so that in some cases, there are no allowed rotations. Then the solution is unique.

In many cases, however, some rotations are possible without violating the non-negativity constraints. Then there are an infinite number of possible solutions to the factor analytic problem. In practice, there are several points to be considered: (1) Is the solution unique or not? (2) If not unique, what is the extent of the domain of possible solutions and (3) Can one solution be found that appears more plausible than the others. These aspects of the factor analysis problem are the focus of the present work.

It is well known that if a sufficient number of elements of **G** (or **F**) are a priori know to be zero, then there is no rotational ambiguity in the solution [6]. Then the problem is easily solved by simply fixing those elements to be zero before starting the least squares fit. One might also assume that a unique result is obtained if a sufficient number of elements are fixed to zero in **G** and **F**, taken together. The authors are not aware of any published analysis of uniqueness in the case when zeros are enforced both among scores and among loadings.

In practice, one may have good reason to believe that there are zero values in **G** (or/and in **F**), without having any a priori information about the locations of those zero elements. Then "fixing to zero", as mentioned in the preceding paragraph, is not applicable. Then non-negativity constraints may be applied instead, in the expectation that some elements of **G** (or/and of **F**) will become zero in the fitting process.

1.1.1. Terminology

The term *rotation* is traditional. It would be more correct to use the term *linear transformation*. However, it is customary to use the word rotation. Sometimes the terms *orthogonal rotation* and *oblique rotation* are used in order to distinguish between cases where the matrix **T** is and is not orthogonal. In this work, orthogonality of **T** is not of interest. Thus, the word rotation is used in the meaning of oblique rotation, i.e. arbitrary non-singular linear transformation.

The word "rotation" is used here in two meanings. The exact meaning, a *pure rotation*, is defined above. The word is also used for *approximate* or *distorted rotations*, where Eq. (4) are only approximately obeyed. The typical reason for distortion is that exact rotation would create some negative factor elements. Distortions keep those elements at the constrained value of zero. When a distorted rotation is performed, the product of the two matrices **G** and **F** changes and thus, Q also changes.

The notation $\mathbf{G} \geq 0$ (or $\mathbf{F} \geq 0$) is used in the meaning that each element of the matrix **G** (or **F**) fulfills the inequality.

2. Elementary rotations

2.1. Definition of an elementary rotation

As discussed by Paatero and Tapper [7], a specific rotation may be understood as consisting of a sequence of many elementary rotations. If no elementary rotations are possible without creating negative values, then it can be concluded that the solution is unique. In order to discuss elementary rotations, the following index conventions are assumed: s and u denote some chosen factor indices so that $1 \leq s \leq p$, $1 \leq u \leq p$, $s \neq u$. The symbol v denotes all index values 1,. . .,p except s. Similarly, w denotes all index values 1,. . .,p except u. The index i represents all values from 1 to m. Similarly, j represents all values from 1 to n. The elementary rotation between factors s and u is represented by equations:

$$\begin{aligned} \bar{g}_{is} &= g_{is} + r g_{iu}, \quad \bar{g}_{iv} = g_{iv} \\ \bar{f}_{uj} &= f_{uj} - r f_{sj}, \quad \bar{f}_{wj} = f_{wj} \end{aligned} . \qquad (5)$$

The matrix **T** corresponding to Eq. (5) is almost equal to a unit matrix of dimension $p \times p$. The only difference is that the off-diagonal element $t_{us} = r$. One can verify that Eq. (5) really represent a rotation by computing the matrix product $\bar{\mathbf{G}}\bar{\mathbf{F}}$ based on Eq. (5) and verifying that $\bar{\mathbf{G}}\bar{\mathbf{F}} = \mathbf{GF}$. In this case, the inverse matrix $\mathbf{T}^{-1}$ is also surprisingly simple. It is like matrix **T**, except that its only non-zero off-diagonal element (us) is of opposite sign in comparison to the same element in **T**.

Eq. (5) may be conceptualized as follows: The values in column u of **G**, multiplied by r, are added to column s of **G**. Simultaneously, the values in row s of **F**, multiplied by $\mathbf{r}$, are subtracted from row u of **F**.

It should be stressed that elementary rotations are basically a tool for understanding rotations. For actually performing the rotations, elementary rotations are useful when searching for a good rotation "by hand", so that one performs incremental rotations upon the factors and observes the results before trying the next incremental rotation. In automatic computations with programs PMF and ME, all rotations are computed in a global way, without recourse to elementary rotations.

2.2. Elementary rotations applied to the correct solution

Assume that the correct solution, corresponding to the physical components of the situation, is represented by the factor matrices **G** and **F**. Examine what happens when an elementary rotation with coefficient r is applied to this solution. If r is positive, then all changes induced in matrix **G** are in the positive direction. No element in **G** can decrease. In contrast, all changes in **F** are in the negative direction, and no element in **F** may increase. If, on the other hand, r is negative, then the behavior of **G** and **F** is the reversed: Elements in **G** either decrease or stay constant, while elements in **F** either increase or stay constant.

If all elements in the true factor matrices, **G** and **F**, are strictly positive, then all rotations are possible in the following sense: if arbitrary factors s and u are chosen for rotation, then there are limits $r_- < 0$, $r_+ > 0$ that define the domain of allowed rotations. Whenever $r_- \leq r \leq r_+$, the rotation does not produce any negative elements in either of the matrices **G** or **F**. If r equals r_- or r_+, then one element in one of the rotated factor matrices becomes exactly zero. It is seen that there are $p^2 - p$ possible types of elementary rotations. Furthermore, regarding each one of these possible rotations, the true solution is situated in an inner point of the domain of possible rotations. The number of possible elementary rotations is the same as in ordinary unconstrained PCA. However, some of their rotational domains may be rather small, depending on the numerical values of the factor elements. If the domain is small for a certain rotation, then the rotation in question may in practice be insignificant and might perhaps be ignored.

Next, consider the situation when all elements in the true factor matrix **G** are strictly positive, while there are a number of zeroes among the elements of **F**. Rotations with negative parameter values r will increase elements of **F** and decrease those of **G**. All such rotations are possible (with sufficiently small r) because each element of **G** may decrease at least a little before hitting against zero. Rotations with positive values of r offer a different picture. Choose some factors s and u for rotation. If there is a value j such that $f_{uj} = 0, f_{sj} > 0$, then the corresponding rotated element becomes negative: $\bar{f}_{uj} < 0$. Then the rotation (s,u) with $r > 0$ is not possible. Considering this rotation, the true solution is situated at the extreme end of the rotational domain.

If the number of zero entries in **F** is large enough, then all possible rotations with arbitrary indices (s,u) are impossible with positive values of r. Then the true solution is situated at the extreme end of each of the rotational domains.

Third, consider the situation when there are a number of zeroes among the elements of both **G** and **F**. Choose some factors s and u for rotation. If there is a value j such that $f_{uj} = 0$, $f_{sj} > 0$, then the rotation (s,u) is not possible with $r > 0$. Analogously, if there is a value i such that $g_{is} = 0$, $g_{iu} > 0$, then the rotation (s,u) is not possible with $r < 0$. It is seen that if there are a sufficient number of zero values in both the **F** and **G** matrices, then non-negativity causes the solution to be rotationally unique: there are no possible rotations.

2.3. Example: the true solution at the end of the rotational domain

Spectroscopic problems offer examples where the true solution can be at the extreme end of the rotationally accessible domain. Assume that the columns of the true **G** consist of spectral peaks whose shapes are Gaussian with identical widths. Then subtracting any one of the columns from another will produce negative values regardless of the size of r. Then all rotations with a negative r are impossible.

3. Global control of the rotations of the computed solution

3.1. The true solution vs. the computed solution

In the preceding discussion, the conditions under which the true solution may be rotated were presented. In practice, the true solution is unknown, and only the best computed solution is known. The same logic can be applied to the computed solution. If it possesses a sufficient number of zero values, it cannot be rotated at all. Then a unique solution has been obtained.

On the other hand, it may happen that one believes that the true solution is at the extreme end of the rotational domains, while the computed solution is seen to be in the middle of (most of) the rotational domains. Then it is clear that there is rotational ambiguity and that the computed solution is not the correct one. From the reasoning presented above, one knows that the *true* solution cannot be rotated with positive r, say. Then one might try to rotate the *computed* solution with positive r, hoping that the rotations would be transformed up to the true solution but not past it.

3.2. The rotational parameter ϕ in the program PMF2

Up to this point, the discussion has been valid for all possible means of computing two-way factor analysis, including PCA and manual methods of finding "good" rotations. Now the rotational machinery of PMF2 is discussed.

The rotations computed by PMF2 are influenced by two special features. First, non-negativity is implemented with logarithmic penalty functions. These functions cause the solutions to stay away from all zero values. One could say that PMF2 attempts to compute a solution that is in the middle of all rotational domains. Second, a non-zero value of the user-specified rotational parameter ϕ, called FPEAK in the documentation of PMF, tries to impose rotations to the emerging solutions throughout the iteration sequence. Positive values of ϕ try to rotate using positive coefficients r in Eq. (5). Similarly, a negative ϕ tries to rotate with negative coefficients r.

If the true solution is not at the extreme end of all rotational domains, then rotating with ϕ is not likely to find the true solution. In other words, if there are not enough of zero values in either the **G** and **F** matrices, then rotating with ϕ has no theoretical justification. In this case, there is no best value of ϕ to use. Unless specific patterns of zero valued elements in either **G** or **F** or both are known a priori, there is no theoretical basis for choosing a specific value of ϕ. The domain of rotations can be explored by choosing a series of specific values and examining the results. However, any specific choice of the ϕ value will have to be based on the judgment of investigator based on their individual knowledge of the system under study.

However, using empirical estimations about suitable values of ϕ can potentially cause problems. An investigator may assume that the new measurement is similar to previous measurements, and thus, it can be successfully analyzed by setting ϕ equal to some specific value that made sense in the prior application. In environmental monitoring and enforcement, unusual situations do occur. Sometimes these unusual situations are the most critical ones where most unpredictable circumstances may occur. The patterns of factors may be different from the familiar ones. Thus, the prior information about ϕ cannot be extended to the analysis of unusual data.

When analyzing environmental data, the most important use of the parameter ϕ is to quickly demonstrate how much rotational ambiguity exists in the computed factors. By computing with both positive and negative values of ϕ the range of possible solutions can be bracketed. Using ϕ for pinpointing one solution within this range is not possible except in rare special occasions.

What happens if the original data matrix **X** is transposed and the transposed matrix $\mathbf{X}^{\mathrm{T}}$ is analyzed instead of **X**? Then the roles of the factor matrices **G** and **F** are interchanged: the original contents of **G** appear now in **F**, and vice versa. If the original problem is solved successfully with a positive ϕ, then the transposed problem requires a negative ϕ, and vice versa. It is seen that when reporting about usage of ϕ, it is not sufficient to say that "problems of type xxx require rotation with a positive ϕ". In addition, one must state the arrangement of the problem, i.e. what are the roles of the rows and columns of **X**. In source apportionment problems, one must specify whether one column of **X** represents the concentration of one element in all samples or the concentrations of all elements in one sample. Equivalently, one might say that "the rows of **F** represent emission profiles of sources" or "the rows of **F** represent time sequences of emissions by different sources".

3.3. Rotations in other multilinear programs

Managing rotations in PMF2 is based on the special structure of the program. The rotational matrix, **T**, is explicitly manipulated in the program. In three-way factor analytic program, PMF3, there is no corresponding mechanism for controlling rotations.

Paatero [3] describes a generic technique for controlling rotations. This technique is based on an enhanced form of the object function. The technique is generic in the sense that it may be applied in any factor analytic algorithm that is explicitly based on minimizing some form of a quadratic object function. When using the multilinear engine program, ME-2 [8], for solving non-negative factor analytic problems, the technique has been used successfully in a modified form. The method will be described for obtaining rotations with negative r. For the opposite case $r > 0$, the roles of **G** and **F** are to be interchanged.

The following additional terms are included in the object function Q:

$$Q^{\mathrm{n}} = \sum_{h=1}^{p} \left(1 - \sum_{j=1}^{n} f_{hj}^2\right)^2 \Big/ \alpha^2 \tag{6}$$

$$Q^{\mathrm{p}} = \beta^2 \left(\sum_{h=1}^{p} \sum_{i=1}^{m} g_{ih}\right)^2. \tag{7}$$

The term Q^{n} defined by Eq. (6) attempts to normalize the rows of **F** to unit norm (sum of squares of elements of each row is driven towards unity). The parameter α should be chosen small enough so that in the computed solution, the norms of rows of **F** deviate from unity at most by a small value (e.g. 0.01). The term Q^{p} defined by Eq. (7) attempts to pull the sum of all elements of **G** towards zero. A tedious analysis shows that this attempt leads to rotations in the direction of negative r. The parameter, β^2, corresponds to negative values of ϕ in PMF2. By increasing β^2, a stronger rotational influence is applied to the solution. For simplicity, rotations are discussed in the following section from the viewpoint of PMF2 and ϕ. The same reasoning applies to ME-2 and the parameter β^2, however. Positive values of ϕ correspond to such Eqs. (6) and (7) where **G** and **F** have been interchanged.

3.4. Flexibility of the computed factors

An apparent contradiction may be observed when computing the same problem with zero and non-zero ϕ values. With no rotational influence, the solution may appear unique. There may be small non-zero factor values that appear to prevent all rotations. However, it has to be remembered that the rotation becomes part of the solution to the problem. Thus, when the solution is recomputed with a non-zero ϕ, a considerable amount of rotation may be observed. The explanation is as follows: in the "academic" discussion of the rotations, the factors are assumed rigid in the sense that the product of the factor matrices must not change at all. In real life, the factors are "alive" or flexible: any changes in factors are acceptable as long as the Q value does not grow "too much." Under the influence of ϕ, the program will accept a slightly worse fit in order to minimize the modified object function. In that way, a distorted rotation is performed. This worse fit is caused because a number of factor elements need to deviate from the negative values. The question of how much increase of Q is "too much" is discussed below.

The experience accumulated when rotating with ϕ may be summarized as follows: the first phase is when ϕ is initially increased from zero. Q increases slowly during this first phase. If there is rotational freedom in the original result, then this freedom is "consumed" during the first phase. The solution rotates as far as it may easily go. If the original solution is rotationally fixed, then the solution hardly changes at all. In the second phase, when ϕ is increased further, Q increases steeply. The factor shapes change clearly because now the factors have become distorted. The changes of factor elements are no longer fully described by Eq. (5). It appears that useful results are often obtained with the ϕ that corresponded to the end of the first phase. Further experience is needed about choosing useful values of ϕ.

3.5. Caveats

The intuitive notion of elementary rotations gives a feeling about the possibility of rotations. However, considering elementary rotations alone does not give rigorous conditions about the uniqueness of a solution. Such non-unique solutions exist where no single elementary rotation is possible. A combination of two (or more) simultaneous elementary rotations may still lead to a different allowed (non-negative) solution even in cases where all single rotations are impossible. This warning does not affect practical computations, e.g. when pulling selected factor elements to zero, because all fitting is performed in a global way, not through elementary rotations.

Rotations driven by ϕ do not necessarily lead to a "best" rotated solution. The iteration may end up in situations where no continuation is possible by using only rotations with the desired sign of r. Continuation would require a combined rotation, composed of elementary rotations having different signs of r. There is no guarantee that any value of ϕ will force such a combined rotation. Thus, there is also no guarantee that the true solution is contained in the domain covered by the rotations driven by different values of ϕ.

Usually, factor analytic least squares computations are started from pseudorandom initial values. When studying rotations, the situation is different. If started from same initial values, iterations may end up in different local minima depending on what rotational parameters are used. Also, identities of factors may change, i.e. different sources appear with different indices h. The effect of any rotational influence may be entirely masked by the different features of the different local solutions. When studying rotations, the following scheme is suggested. First, run multiple pseudorandom starts without rotational parameters, or with a weak (= safe) rotational forcing. Inspect the results and choose one or a few most important solutions for use as starting points in further study. Set up a run where rotations are forced. Use the chosen solution as the starting point and input the initial factor values from a saved file rather than generating new pseudorandom values. Repeat the experiment with different rotational parameter values, and each time use the same chosen solution as the starting point. Possibly, repeat the same exercise with another chosen solution as the starting point. In this way, it will be possible to compare effects caused by slightly different rotational forcings.

4. Techniques for estimating the rotational ambiguity

4.1. What is meant by rotational ambiguity?

Rotational ambiguity, as a property of a solution ($\mathbf{G}$,$\mathbf{F}$), may be understood as the collection T of all "allowed" rotational transformations $\mathbf{T}$ (both elementary and non-elementary rotations included) that may be applied to ($\mathbf{G}$,$\mathbf{F}$). The formal interpretation of this definition would be:

$$T = \{\mathbf{T} \mid \mathbf{GT} \geq 0,\ \mathbf{T}^{-1}\mathbf{F} \geq 0\}. \tag{8}$$

However, this formal definition ignores the flexibility of factors when attempting a rotation. Instead, the following intuitive definition is more realistic:

$$T = \{\mathbf{T} \mid \bar{\bar{\mathbf{G}}} \approx \mathbf{GT},\ \bar{\bar{\mathbf{F}}} \approx \mathbf{T}^{-1}\mathbf{F},\ \bar{\bar{\mathbf{G}}} \geq 0,\ \bar{\bar{\mathbf{F}}} \geq 0,\ Q(\bar{\bar{\mathbf{G}}}, \bar{\bar{\mathbf{F}}}) \leq Q^{\max}\} \tag{9}$$

where $Q^{\max}$ denotes the largest acceptable value for Q (see below).

It does not seem possible to characterize the ambiguity by one single numerical value. Different tools for characterizing the family of acceptable rotations $\mathbf{T}$ are needed. One possibility is to explore how small or how large a few chosen individual elements of $\mathbf{G}$ and/or $\mathbf{F}$ may become.

4.2. Assessing the increase of Q when a rotation is attempted

Throughout this work, rotations are considered acceptable if they do not increase Q "too much." Full understanding of how much is "too much" needs to be provided. Devising reliable criteria is complicated by the fact that the error structure of environmental measurements is complicated. Some error sources may be common to many data points, and hence statistical techniques based on the assumed statistical independence of errors of data points are not applicable. The following qualitative guidelines have been used successfully in practical work. It is assumed that correct error estimates have been used in the analysis. The notation Q^{m} means the portion of Q that arises from the main or data-fitting equations (penalty-defining equations are excluded).

If the model is correct, the expected value of Q^{m} is approximately equal to the number of data values minus the number of essential free parameters fitted to the data. This value is often called degrees of freedom v. For factor analysis, this gives:

$$E(Q^{\mathrm{m}}) \approx v = nm - p(n+m) \tag{10}$$

i.e. the size of $\mathbf{X}$ minus the number of unknown factor elements. (Because of free rotations, the number of essential free parameters is usually smaller than the number of factor elements. However, this problem is ignored.) Consider a rotation that meets resistance because a number of factor elements are driven against the limit of zero. Then Q^{m} increases because the shapes of factors must change more than what corresponds to a pure rotation. On the other hand, the factor elements that are forced to be approximately zero are not free variables any more. Their values cannot be adjusted in order to optimize the fit. Thus, one may argue that the number of free parameters should not include such factor elements that are approximately zero. It follows that the expected value of Q^{m} increases by the number of (near) zero entries that are introduced in the factor matrices because of the rotation. The practical rule is as follows: "A rotation that introduces k (near-)zeros in factor matrices is certainly allowable if the accompanying increase of Q^{m} is less than k units."

No rule is known that would say when a rotation is certainly not allowable. In practice, increases of Q^{m} by tens of percent have been considered forbidden. Between these two extremes, there is the gray area where reliable criteria are not yet known. When reporting results where rotations in the gray area are considered, one should probably report the observed increase of Q^{m} so that the reader may form an opinion about the situation.

4.3. Approaching rotational ambiguity through trial-and-error rotations

Detailed information about the rotational ambiguity may be obtained by trial-and-error rotations. Then one repeatedly influences the solution by any of the available methods (e.g. by pulling chosen factor elements towards zero) and observes how much the Q^{m} value increases from its lowest value. By comparing the change induced in results and the change of Q^{m}, one obtains information about the uniqueness of the solution. If the increase of Q^{m} is allowable, then the rotated factors represent another viable solution.

Different rotations interact with each other. Thus, it is not enough to determine which individual rotations have large uncertainties. A complete understanding of the situation requires that one also studies to what extent those rotations may be performed simultaneously with each other.

4.4. Program PMF2 and the matrix of rotational uncertainties

Program PMF2 implements non-negativity constraints by means of logarithmic penalty functions, included in the object function Q to be minimized. Whenever an element g_{ih} (or f_{hj}) is about to become negative, then the corresponding term in the object function, $-\ln(g_{ih})$, approaches plus infinity and causes an increase of Q. The elements never become exactly zero. By controlling the strength of the penalty terms, the user may allow arbitrarily close approach to zero but the exact value of zero is never reached.

At each iteration step, PMF2 computes and executes a rotation matrix $\mathbf{T}$ as a linearized least squares fit that minimizes the penalty functions (see Paatero [2]). The unknowns in this fit are all the non-diagonal elements of $\mathbf{T}$. If this rotation is computed at the point of convergence, then the identity matrix is obtained, i.e. all non-diagonal elements of $\mathbf{T}$ are zero. In addition to matrix $\mathbf{T}$, the least squares fit produces a covariance matrix for the estimated unknowns, i.e. for all non-diagonal elements of $\mathbf{T}$. This matrix is of dimensions $(p^2-p)\times(p^2-p)$. The square roots of the diagonal elements of this matrix represent the standard deviations of the computed elements of $\mathbf{T}$. The program PMF2 provides these standard deviation values as a matrix called "rotmat". In this presentation, matrix "rotmat" is denoted by $\mathbf{\Gamma}$. Elements of $\mathbf{\Gamma}$ represent error estimates of elements of the rotation matrix $\mathbf{T}$:

$$\Gamma_{ij} = \mathrm{stddev}(t_{ij}). \tag{11}$$

The significance of elements of $\mathbf{\Gamma}$ is discussed in the PMF User's Guide, Part 2 [9]. Roughly, each value $\mathbf{\Gamma}_{us}$ indicates how large the parameter r may be at most in the corresponding Eq. (5) without increasing the penalty function more than by one unit ($\Delta(Q^{\mathrm{penalty}}) \leq 1.0$). This estimate takes into account all possible rotations, not just elementary ones.

At this time, no quantitative evaluation of $\mathbf{\Gamma}$ has been performed. It is only reasonable to consider it as a qualitative or heuristic tool. Each "large" element in

Γ indicates that there is a free rotation between the indicated factors. In contrast, a "small" valued element of **Γ** means that there is no rotational freedom between the indicated factors. The elements of **Γ** describe the total effect of all possible rotations, elementary and non-elementary. This result is guaranteed by the way **Γ** is formed: its elements are not formed one-by-one by trying individual elementary rotations. Instead, all elements of **Γ** arise from one global least squares fit where all of the non-diagonal elements of **T** are determined together.

The rotational status of the solution may be influenced by pulling some factor elements towards zero or by using a non-zero rotational parameter ϕ. Such influencing will decrease the remaining rotational uncertainty in the new result. Hence, the elements of **Γ** will be smaller for the new solution. This result should not be construed as a proof that the new solution is "better." The smaller elements of **Γ** simply indicate that there is less room for further rotations because the user has fixed some parts of the solution.

In addition to the joint rotational uncertainty of **G** and **F**, there is also individual uncertainty in **G** and **F** (see next section). It is not known how to combine these different estimates into a single description of uncertainty.

4.5. Using the Jacobian or Hessian matrix for estimating uncertainties

Paatero [8] discussed the use of the Jacobian or Hessian matrix for estimating uncertainties of computed factor elements. The approach appears promising for small and medium-sized problems where the data value errors are relatively small. Also, by using the Hessian approach, a single set of estimates is obtained that describes both rotational and regression-like uncertainty of all factor elements. This outcome is in contrast to the approach of PMF2, where a complete uncertainty estimation requires three separate steps: (1) Estimating uncertainty of **G**, when **F** is assumed known, (2) similarly, estimating uncertainty of **F**, when **G** is assumed known, and (3) estimating the joint rotational uncertainty of **G** and **F**, as described above. However, there is no documented record of the application of this approach to date. Further discussion of the approach has to be postponed until more experience is gained.

5. Other techniques for controlling rotations

As a tenet of our scientific experience, there is an implicit assumption that "A unique answer exists to each question. This unique answer will be obtained if you use the right techniques." In the spirit of this assumption, one would hope that there would be some technique for finding "correct" rotations. Unfortunately, it seems that the rotational problem is unsolvable unless some additional information is available. Thus, one has to look for possible sources of additional or a priori information.

5.1. Pulling selected factor elements towards zero

When analyzing Hong Kong aerosol, Lee et al. [10] found sulfate in almost all factors. Such a result did not appear plausible. The concentration of sulfate was then forced toward zero in a number of factors in a follow-up PMF2 run. The increase of Q^{m} did not appear significant and more plausible factors resulted from the pulled-down computations.

The analysis of Hong Kong data offers a simple example of forcing toward zero with only one compound subjected to pulling. It would be possible to use such forcing on several compounds, provided that there is reliable information about compounds that are not emitted by certain sources. In addition, it might also be possible to apply forcing to specific time series factor elements. Sometimes information about weather patters might reveal that a certain source cannot affect the receptor site during certain time intervals or it was not functioning (e.g. plant was on strike). Then the time factor elements corresponding to such intervals could be forced towards zero.

If pulling-down is used in a reported study, then it is essential that such forcing is reported. Of course, such a priori information that justifies the use of pulling must also be reported. In this way, the reader will be provided with all of the pertinent information that is needed for replicating the analysis or for judging the validity of the analysis. Also, at a later time, it may become known that some part of the assumed a priori information was in fact not true. Then proper reporting of the details of the analysis helps in deciding what portion of the results remain valid and what should be discarded.

5.2. Target shapes—half-way between FA and CMB

Factor analysis (PCA, FA) and Chemical Mass Balance (CMB) are two extreme forms of emission analysis. For CMB, see Cooper et al. [11]. In factor analysis, nothing is assumed about the composition profiles of sources (except for the assumption that profiles do not change with time). In CMB, it is assumed that all profiles are exactly known in advance.

As an intermediate between FA and CMB, a method called Target Transformation Factor Analysis (TTFA) has been used [12]. In TTFA, the user specifies likely target shapes for the composition factors. The algorithm attempts to rotate the computed solution so that the target shapes are reproduced as well as possible. Although TTFA has been successful in many practical problems, it suffers from the fact that rotations are performed a posteriori, after choosing the subspace with an eigenanalysis.

Both of the extremes (FA, CMB) have certain merits. However, by using new solution tools such as PMF2 or ME-2, it is possible to set up analyses that are between these extremes where something is assumed known about emission profiles but additional information is to be determined during the analysis. Such analyses might be called with the generic name of "Target factor analysis". It is believed that in situations where one might consider using either CMB or FA, a technique based on target shapes would in fact be optimal, better than either one of the extreme alternatives.

Simple forms of target factors may be implemented by using the program PMF2. However, setting up such runs is tricky and error-prone because PMF2 was not originally designed for these kinds of analyses. The program ME-2 is the better tool for setting up target factors. The technique will be presented without regard to the exact form of setting up the runs.

5.2.1. A hybrid setup between CMB and FA

A hybrid analysis is conceptualized in the following way: a number of the composition factors (= rows of factor matrix **F**) are specified as completely known or "fixed". This is in spirit of CMB. The remaining composition factors are assumed to be completely unknown, in the spirit of PCA or FA. In this way, the best of CMB and of FA could be combined if a few of the sources are well known in advance. This kind of analysis has not been attempted so far.

5.2.2. Target shapes: approximately known ratios of concentrations

Equations of the type:

$$af_{js} - bf_{ju} = 0 \tag{12}$$

are easily implemented in the program ME-2. Such an equation specifies that the ratio of the two concentrations f_{js} and f_{ju} has to be approximately b/a. The error estimate assigned to the equation specifies how closely the factor values must match the specified ratio. Sometimes the absolute level of certain concentrations is not known although the ratios of a few of them are well known. Then Eq. (12) is preferable to techniques presented below.

5.2.3. Target shapes: exactly known concentrations

If some concentrations of a source are reliably known, then the corresponding factor elements may be specified as "fixed". Alternatively, upper and lower limits may be specified for such factor elements so that a narrow interval of variation is left for the values in question. Note that the normalization of the factor in question becomes defined by such specifications. No additional normalization should be specified for such a factor. This arrangement may lead to a fast convergence of the run.

5.2.4. Target shapes: approximately known concentrations

If elements of a composition profile are approximately known, then equations should be specified with suitable error estimates that reflect the uncertainty of the information. The known (or assumed) values of factor elements f_{js} are represented by a matrix φ_{js}. Equations of the type $\varphi_{js}=f_{js}$ are inserted in the model for some indices j and s. Error estimates corresponding to the uncertainty of the information are specified for the equations. If there is no information available for certain concentrations, then either the corresponding equations are omitted from the model or else “infinitely large” error estimates are specified for those equations.

5.3. Should one use a priori information for checking the computed factors or for computing the factors?

Such information can only be used once and using it for both tasks is a circular problem. With customary techniques, one has used a priori information for controlling the quality of the computed analysis. If the computed factors were in conflict with the a priori known facts then the analysis was rejected. If a priori information is used for guiding the analysis towards a meaningful rotation, then this possibility of a posteriori checks is lost. Perhaps a trade-off may be found. Some part of the available a priori information can be reserved for a posteriori checking while the rest is used for setting up auxiliary a priori equations, e.g. Eq. (12).

*5.4. Using the “edges” in the p-dimensional space of rows of matrix **G***

Henry [13] has suggested a technique based on patterns of data points in the p-dimensional space of rows of matrix **G**. These patterns are called *edges*. Simplified, the reasoning goes as follows.

Each data point corresponds to a row in matrix **G**, and hence to a point in a p-dimensional linear space of factors. Non-negativity means that the points are confined to the first octant of the space. Each different rotation corresponds to a different position of the coordinate planes or subspaces in the space. The data points are considered to be in fixed positions. It is assumed that there are one or several sources that are not present in a sufficient fraction of the data points. The points where one source is absent lie in a subspace of dimension $p-1$. In the terminology of Henry, such a subspace is called an edge. Henry has implemented an algorithm UNMIX that attempts to rotate so that one of the coordinate subspaces becomes identical with an edge subspace. If this succeeds, then it is highly probable that the factor in question corresponds to a real source. Often, it is possible to rotate so that several edge subspaces become identical with coordinate subspaces, thus identifying several factors. The use of edges is under development for the multilinear engine.

6. Summary

Elementary rotations were introduced. These rotations serve as an aid in visualizing what happens in a rotation. Also, they allow a quick qualitative way of inspecting the results and deciding if rotational ambiguity is or is not present.

Different techniques for influencing the rotational status of computed factors have been discussed. It is emphasized that these techniques are not statistical miracles. Instead, the techniques are based on assumptions about true factors, and in particular, on assumptions regarding the zero elements in scores and/or loadings. If such assumptions are not justified, then there is also no justification for any of the rotational techniques.

The problem of diagnosing the rotational ambiguity was also discussed. It was seen that the two problems (influencing and diagnosing rotations) are intertwined: the most general method for diagnosis is to try to cause rotations. If rotations do happen, ambiguity is present. Matrix-based methods of diagnosis were also discussed. One such method is based on the internal structure of the program PMF2 and not easily transportable to other environments. Another method, based on the Jacobian or Hessian matrices of the fit, is theoretically superior in the small-error case where a linear approximation of the bilinear model is good enough. However, the Jacobian and Hessian matrices become impractical to handle if the total number of scores and loadings is large (more than, say, 2000).

References

[1] R.C. Henry, Current factor analysis models are ill-posed, Atmos. Environ. 21 (1987) 1815–1820.

[2] L.A. Currie, R.W. Gerlach, C.W. Lewis, W.D. Balfour, J.A. Cooper, S.L. Dattner, R.T. DeCesar, G.E. Gordon, S.L. Heisler, P.K. Hopke, J.J. Shah, G.D. Thurston, H.J. Williamson, Interlaboratory comparison of source apportionment procedures: results for simulated data sets, Atmos. Environ. 18 (1984) 1517–1537.

[3] P. Paatero, Least squares formulation of robust non-negative factor analysis, Chemom. Intell. Lab. Syst. 37 (1997) 23–35.

[4] E.S. Park, R.C. Henry, C.H. Spiegelman, Estimating the number of factors to include a high-dimensional multivariate bilinear model, Commun. Stat.-Simula 29 (2000) 723–746.

[5] J.-H. Wang, P.K. Hopke, T.M. Hancewicz, S. Zhang, Sup-

pressing unnecessary factors in factor analysis, manuscript in preparation (2001).

[6] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 2nd edn., Wiley, 1984.

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

[8] P. Paatero, The multilinear engine—a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model, J. Comput. Graphical Stat. 8 (1999) 854–888.

[9] P. Paatero, PMF User's Guide, Parts 1 and 2, University of Helsinki, Department of Physics (2000). Available from: ftp://rock.helsinki.fi/pub/misc/pmf/.

[10] E. Lee, C.K. Chan, P. Paatero, Application of Positive Matrix Factorization in source apportionment of particulate pollutants in Hong Kong, Atmos. Environ. 33 (1999) 3201–3212.

[11] J.A. Cooper, J.G. Watson, J.J. Huntzicker, The effective variance weighting for least squares calculations applied to the mass balance receptor model, Atmos. Environ. 18 (1984) 1347–1355.

[12] P.K. Hopke, Target Transformation Factor Analysis as an aerosol mass apportionment method: a review and sensitivity analysis, Atmos. Environ. 22 (1988) 1777–1792.

[13] R.C. Henry, Receptor model applied to patterns in space (RMAPS) part I—model description, J. Air Waste Manage. Assoc. 47 (1997) 216–219.