Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 490 (2003) 277–289
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Discarding or downweighting high-noise variables in factor analytic models

Pentti Paatero [a], Philip K. Hopke [b,*]

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

Received 7 October 2002; received in revised form 12 December 2002; accepted 17 December 2002

Abstract

This work examines the factor analysis of matrices where the proportion of signal and noise is very different in different columns (variables). Such matrices often occur when measuring elemental concentrations in environmental samples. In the strongest variables, the error level may be a few percent. For the weakest variables, the data may consist almost entirely of noise. This paper demonstrates that the proper scaling of weak variables is critical. It is found that if a few weak variables are scaled to too high a weight in the analysis, the errors in computed factors would grow, possibly obscuring the weakest factor(s) by the increased noise level. The mathematical explanation of this phenomenon is explored by means of Givens rotations. It is shown that the customary form of principal component analysis (PCA), based on autoscaling the original data, is generally very ineffective because the scaling of weak variables becomes much too high. Practical advice is given for dealing with noisy data in both PCA and positive matrix factorization (PMF).
© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Principal component analysis; Positive matrix factorization; Signal-to-noise; Scaling of variables; Autoscaling; Weak variables; Givens rotations

1. Introduction

The question of accepting or rejecting individual components in principal component analysis (PCA) has been studied by many authors over several decades. In comparison, the question of accepting or rejecting individual variables has received little attention. This question becomes acute whenever the noise content in a specific variable(s) is much higher than exists in other variables. Such a situation may commonly occur in environmental studies. For example, the noise level in some variables may be well below 10% of the signal. While in other variables, there may be more noise than signal, or possibly no signal at all. A few authors have remarked that it is crucial to properly select the chemical elements to be included in a factor analytic study of aerosol samples [1,2]. In the present work, the analysis of noisy variables is studied both analytically and by means of simulations.

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

0003-2670/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0003-2670(02)01643-4

Nomenclature

diag($\boldsymbol{d}$)	a diagonal matrix whose diagonal elements come from vector $\boldsymbol{d}$
$\boldsymbol{E}^0$	the original error matrix with elements e_{ij} from $N(\boldsymbol{0}, \boldsymbol{I})$ distribution
p	index for columns of $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$, $p = 1, \ldots, P$
P	the rank of $\boldsymbol{X}^0$
PCA	principal component analysis
PMF	positive matrix factorization
s_{pp}	the pth singular value of any version of the data matrix $\boldsymbol{X}$
S/N	signal-to-noise ratio
SVD	singular value decomposition, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$
x_{ij}	element i, j of the matrix $\boldsymbol{X}$
$\boldsymbol{x}_j^0$	column j of the matrix $\boldsymbol{X}^0$
$\boldsymbol{X}$	a data matrix of dimensions $m \times n$
$\boldsymbol{X}^0$	the true version of $\boldsymbol{X}$ without random noise, $\boldsymbol{X}^0 = \boldsymbol{U}^0\boldsymbol{S}^0\boldsymbol{V}^{0\mathrm{T}}$
XRF	X-ray fluorescence analysis

Greek letters

σ_{pp}	the pth diagonal element of $\boldsymbol{S}^0$, i.e. the pth singular value of $\boldsymbol{X}^0$
σ_{PP}	the smallest non-zero singular value of $\boldsymbol{X}^0$

1.1. Weak and bad variables

A variable (column of matrix $\boldsymbol{X}$) will be called "*a weak variable*" if it contains signal and noise in comparable amounts. Similarly, variables containing much more noise than signal are termed "bad variables". This work will explore how to deal with weak and bad variables when performing factor analysis with the techniques PCA and positive matrix factorization (PMF) [3]. Practical suggestions will be given for treating both weak and bad variables.

It is problematic to give precise definitions of the terms weak and bad variables. On the one hand, currently there is not much practical experience about the suggested techniques. On the other hand, the shapes of distributions will be different for different variables. This should be taken into account when defining what *weak* or *bad* means. Also, the definition will depend on the representation of smallest concentrations: are they given as measured, or are they censored so that only the detection level (DL) is given instead of such actual values that are "below detection level" (BDL).

The definition of weak variables in the uncensored case is tentatively based on the signal-to-noise ratio (S/N). Denote the signal vector by $\boldsymbol{s}$ and the noise vector by $\boldsymbol{n}$. Then S/N of $\boldsymbol{s} + \boldsymbol{n}$ is defined as $\mathrm{S/N} = \sqrt{\sum s_i^2 / \sum n_i^2}$ (note that in many areas of science, e.g. in electronics and acoustics, the square root is not included in the definition of S/N). A variable is defined to be weak if its S/N is between 0.2 and 2. If $\mathrm{S/N} < 0.2$, then a variable is defined to be bad. As already suggested, these limits are somewhat arbitrary. The intention is that in a weak variable, there are similar amounts of signal and noise, while there is clearly less signal than noise in a bad variable.

For censored data, a tentative definition is based on the number $m_{\mathrm{DL}\,j}$ of BDL values in column j. Denote the (average) detection limit for censored values in column j by δ_j. Then variable j is defined to be weak if

$$0.2 < \frac{\sum_{\{i|x_{ij}>\delta_j\}} x_{ij}}{\delta_j m_{\mathrm{DL}\,j}} < 2 \tag{1.1}$$

An example of this definition: assume that 50% of values x_{ij} are BDL. Denote by X_j the average of those values x_{ij} that are above δ_j. Then Eq. (1.1) reduces to $0.2 < X_j/\delta_j < 2$.

1.2. Notation

In order to denote different versions or variants of a matrix, superscripts are used. Thus, the first diagonal element of the initial version of matrix $\boldsymbol{S}$ will be denoted by s_{11}^0. Subsequent versions of matrix $\boldsymbol{S}$ will be denoted by $\boldsymbol{S}^a$. However, the numerical superscript 2 always denotes the second power of the quantity in question.

The notation $N(a, b)$ denotes a normally distributed (pseudo)random quantity with average a, variance b. Similarly, for a (pseudo)random vector or matrix, $N(\boldsymbol{a}, \boldsymbol{B})$ denotes a random variable having multivariate normal distribution with average $\boldsymbol{a}$, covariance matrix $\boldsymbol{B}$.

2. Analytic study of the effect of noise in PCA

2.1. The analytic tools

Analytic results can only be obtained for rather unrealistic models where data errors are assumed to be homoscedastic, statistically independent, and normally distributed. Thus it is assumed in this section that all data errors come from the $N(0, I)$ distribution (this notation implies both independence and homoscedasticity). The analysis is based on transforming the data matrix so that it is multiplied (from left or/and from right) by orthogonal matrices. Heavy use is made of the following obvious fact.

If elements of a random matrix $\boldsymbol{E}$ are statistically independent and drawn from $N(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{U}$ is an orthogonal matrix of correct dimensions whose elements are independent of elements of $\boldsymbol{E}$, then the elements of the product $\boldsymbol{UE}$ (or $\boldsymbol{EU}$) inherit the same statistical properties as the original elements of $\boldsymbol{E}$: they are statistically independent and have the $N(\boldsymbol{0}, \boldsymbol{I})$ distribution. Quite often, certain elements of $\boldsymbol{E}$ are used when specifying the elements of $\boldsymbol{U}$. These $\boldsymbol{E}$ elements then become "dirty". The descendants of dirty $\boldsymbol{E}$ elements in the product matrix also become dirty in that they do not inherit the statistical properties of the original elements of $\boldsymbol{E}$. However, all of the "clean" elements of the product matrix still inherit the statistical properties, provided they do not descend from any of the dirty elements.

Products of Givens rotations (see Appendix A) are used for developing orthogonal matrices. For a detailed discussion of the application of Givens rotations for diagonalizing a matrix, the reader is referred to Golub and Van Loan [4].

Matrix diagrams will be simplified in the following way: the symbol $(*)$ is used for denoting independent $N(\boldsymbol{0}, \boldsymbol{I})$ random elements. Each occurrence of $(*)$ denotes a different random value. Each occurrence of the underlined or boldfaced $(*)$ denotes a vector consisting of different random values that individually would be denoted by $(*)$. If a matrix element is known or required to be zero, then it is denoted by zero, not by a letter or dot. Matrix diagrams are simplified so that similar elements are denoted simply by dots.

The customary formulation of Givens rotations, as needed for diagonalizing a 2 by 2 matrix, is

$$\begin{vmatrix} u & -t \\ t & u \end{vmatrix} \begin{vmatrix} a & b \\ f & d \end{vmatrix} \begin{vmatrix} c & -s \\ s & c \end{vmatrix} = \begin{vmatrix} z_{11} & 0 \\ 0 & z_{22} \end{vmatrix} \tag{1.2}$$

Here, the original matrix to be diagonalized, consists of the arbitrary elements a, b, f, d. The left and right Givens rotations are determined by u and t, and by c and s, respectively. The rotation matrices are required to be orthogonal, i.e. satisfying $u^2 + t^2 = 1$ and $c^2 + s^2 = 1$. The additional conditions, $|t| \leq |u|$, $|s| \leq |t|$ are imposed in order to select the "smallest possible" rotation.

In this work, the following equation is used instead of Eq. (1.2):

$$uc \begin{vmatrix} 1 & -t \\ t & 1 \end{vmatrix} \begin{vmatrix} a & b \\ f & d \end{vmatrix} \begin{vmatrix} 1 & -s \\ s & 1 \end{vmatrix} = \begin{vmatrix} z_{11} & 0 \\ 0 & z_{22} \end{vmatrix} \tag{1.3}$$

where $u = (1 + t^2)^{-1/2}$, $c = (1 + s^2)^{-1/2}$, $|t| \leq 1$, $|s| \leq 1$.

The rotations in Eq. (1.3) are obtained from $t^2 + pt - 1 = 0$, $s^2 + qs - 1 = 0$, where

$$p = \frac{a^2 - d^2 - f^2 + b^2}{af + bd},$$
$$q = \frac{a^2 - d^2 + f^2 - b^2}{ab + fd} \tag{1.4}$$

As expected, these two equations mirror each other through exchange of b and f.

2.2. Distortion of SVD by noise

In this work, the singular value decomposition (SVD) of a given matrix $\boldsymbol{X}$ of dimensions m by n is defined as

$$\boldsymbol{X} = \boldsymbol{USV}^{\mathrm{T}} \tag{1.5}$$

where the dimensions of $\boldsymbol{U}$, $\boldsymbol{S}$, $\boldsymbol{V}$ are, $m \times m$, $m \times n$, and $n \times n$, respectively. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal. All of the non-diagonal elements of $\boldsymbol{S}$ are zero. The diagonal elements of $\boldsymbol{S}$ are non-negative and ordered in non-increasing order.

The following presentation is based on an example of dimensions $m = 7$, $n = 5$. It is assumed that the true (noise-free) matrix $\boldsymbol{X}^0$ has rank $P = 2$. Then the

SVD of $\boldsymbol{X}^0$ has the following structure:

$$\boldsymbol{X}^0 = \boldsymbol{U}^0\boldsymbol{S}^0\boldsymbol{V}^{0\mathrm{T}} = \begin{vmatrix} u^0_{11} & u^0_{12} & u^0_{13} & \cdots & u^0_{17} \\ u^0_{21} & u^0_{22} & u^0_{23} & \cdots & u^0_{27} \\ u^0_{31} & u^0_{32} & u^0_{33} & \cdots & \cdot \\ u^0_{41} & \cdot & \cdot & \cdots & \cdot \\ u^0_{51} & \cdot & \cdot & \cdots & \cdot \\ u^0_{61} & \cdot & \cdot & \cdots & \cdot \\ u^0_{71} & u^0_{72} & u^0_{73} & \cdots & u^0_{77} \end{vmatrix} \begin{vmatrix} \sigma_{11} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} \begin{vmatrix} \nu^0_{11} & \nu^0_{21} & \cdot & \cdot & \nu^0_{51} \\ \nu^0_{12} & \nu^0_{22} & \cdot & \cdot & \nu^0_{52} \\ \nu^0_{13} & \cdot & \cdot & \cdot & \nu^0_{53} \\ \nu^0_{14} & \cdot & \cdot & \cdot & \nu^0_{54} \\ \nu^0_{15} & \nu^0_{25} & \cdot & \cdot & \nu^0_{55} \end{vmatrix} \tag{1.6}$$

Here, the unusual indexing of ν elements is caused by transposing $\boldsymbol{V}^0$. The two first columns of $\boldsymbol{U}^0$, and the two first rows of $\boldsymbol{V}^{0\mathrm{T}}$, are the principal component vectors of $\boldsymbol{X}^0$. They would be the result of a PCA computed of an error-free matrix. The remaining columns of $\boldsymbol{U}^0$ are arbitrary except for the orthogonality conditions. Their role is to assure that a complete basis set is available for columns of $\boldsymbol{X}^0$. Similarly, the remaining rows of $\boldsymbol{V}^{0\mathrm{T}}$ are arbitrary. In actual computations, these arbitrary values will depend on the software used.

As the next step, the matrices $\boldsymbol{U}^0$ and $\boldsymbol{V}^0$ are kept fixed while an error matrix $\boldsymbol{E}$ is added to $\boldsymbol{X}^0$. It is assumed that the elements of $\boldsymbol{E}$ are $N(\boldsymbol{0}, \boldsymbol{I})$. The error-containing matrix $\boldsymbol{X}$ gets the representation

$$\boldsymbol{X} = \boldsymbol{X}^0 + \boldsymbol{E} = \boldsymbol{U}^0(\boldsymbol{S}^0 + \boldsymbol{R})\boldsymbol{V}^{0\mathrm{T}} = \boldsymbol{U}^0\boldsymbol{S}\boldsymbol{V}^{0\mathrm{T}} \tag{1.7}$$

where $\boldsymbol{R} = \boldsymbol{U}^{0\mathrm{T}}\boldsymbol{E}\boldsymbol{V}^0$. The elements of the random matrix $\boldsymbol{R}$ are different from the elements of $\boldsymbol{E}$ but they inherit their statistical properties. The matrix $\boldsymbol{S} = \boldsymbol{S}^0 + \boldsymbol{R}$ has the following structure:

$$\boldsymbol{S} = \begin{vmatrix} \sigma_{11} + * & * & * & * & * \\ * & \sigma_{22} + * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{vmatrix} \tag{1.8}$$

Here, the asterisk symbols denote independent random values, as already explained. Eq. (1.7) does not represent a SVD of $\boldsymbol{X}$ because $\boldsymbol{S}$ is not diagonal. A SVD can be obtained if $\boldsymbol{S}$ is rotated, by using orthogonal rotations, to be diagonal (obtaining a SVD follows directly from the properties that are needed in order that $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$ be a SVD: $\boldsymbol{U}$ and $\boldsymbol{V}$ must be orthogonal, and $\boldsymbol{S}$ must be diagonal, with non-negative entries ordered in a non-increasing sequence). The rotation is represented by

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{U}^0\boldsymbol{S}\boldsymbol{V}^{0\mathrm{T}} = \boldsymbol{U}^0\boldsymbol{U}^a\boldsymbol{U}^{a\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^a\boldsymbol{V}^{a\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}} \\ &= (\boldsymbol{U}^0\boldsymbol{U}^a)(\boldsymbol{U}^{a\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^a)(\boldsymbol{V}^{a\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}) \end{aligned} \tag{1.9}$$

The last expression in this equation would represent a SVD of $\boldsymbol{X}$ if $\boldsymbol{U}^{a\mathrm{T}}\,\boldsymbol{S}\boldsymbol{V}^a$ would be diagonal. However, in general it is not possible to directly find a single pair of rotations ($\boldsymbol{U}^a$, $\boldsymbol{V}^a$) that diagonalizes $\boldsymbol{S}$. Instead, sequences of rotations are needed, leading eventually to

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{U}^0\boldsymbol{U}^a\cdots\boldsymbol{U}^k)(\boldsymbol{U}^{k\mathrm{T}}\cdots\boldsymbol{U}^{a\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^a\cdots\boldsymbol{V}^l) \\ &\quad\times(\boldsymbol{V}^{l\mathrm{T}}\cdots\boldsymbol{V}^{a\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}) \\ &= (\boldsymbol{U}^0\boldsymbol{U}^k)(\boldsymbol{U}^{K\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^L)(\boldsymbol{V}^{L\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}) \end{aligned} \tag{1.10}$$

where the central bracketed expression is diagonal. In fact, it is not necessary to diagonalize the entire $\boldsymbol{S}$. Such rotations are not significant for the error analyses that do not affect any of the P first columns or rows of $\boldsymbol{S}$. Thus, the error analysis is complete when the diagonalization of $\boldsymbol{S}$ is carried out so far that the upper left block is diagonal, the lower left block is zero, and also the upper right block is zero. The lower right block may be left as it is after the sequence of rotations.

2.2.1. Noise level connected with the elements of rotations

The sizes of non-diagonal elements in the P first columns of the overall rotation matrix $\boldsymbol{U}^K = \boldsymbol{U}^a, \ldots, \boldsymbol{U}^k$ determine how much the "noise" columns $P+1, \ldots, n$ of $\boldsymbol{U}^0$ contaminate the "signal" columns $1, \ldots, P$ of $\boldsymbol{U}^0$ in the product $\boldsymbol{U} = \boldsymbol{U}^0\boldsymbol{U}^K$. By

applying the definition of S/N, as presented earlier, one obtains for column p of $\boldsymbol{U}$

$$\mathrm{S/N} = \sqrt{\frac{u_{pp}^2}{\sum_{i\neq p} u_{ip}^2}} \tag{1.11}$$

where the notations u_{ip} and u_{pp} refer to elements of the rotation matrix $\boldsymbol{U}^K$. The corresponding S/N ratio for the pth column of $\boldsymbol{V}^0\boldsymbol{V}^L$ (i.e. pth row of $\boldsymbol{V}^{L\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}$) is

$$\mathrm{S/N} = \sqrt{\frac{v_{pp}^2}{\sum_{i\neq p} v_{ip}^2}} \tag{1.12}$$

Similar expressions may also be formed for characterizing the individual rotations $\boldsymbol{U}^h$ and $\boldsymbol{V}^h$.

2.2.2. The rotations for introducing zeroes in the first columns of S

In the first stage, off-diagonal elements of the first column of $\boldsymbol{S}$ are zeroed so that rows $m, m-1, \ldots, 2$ of $\boldsymbol{S}$ are rotated with row 1. Each rotation is arranged so that one element of the column becomes zero, resulting in matrix $\boldsymbol{S}^a$ in Eq. (1.13).

$$\boldsymbol{S}^a = \begin{vmatrix} s_{11}^a & s_{12}^a & * & * & * \\ 0 & s_{22}^a & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{vmatrix},$$

$$\boldsymbol{S}^b = \begin{vmatrix} s_{11}^b & 0 & * & * & * \\ \varepsilon & s_{22}^b & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{vmatrix} \tag{1.13}$$

Next, off-diagonal elements of the second column of $\boldsymbol{S}$ are zeroed by rotating the second row of $\boldsymbol{S}$ with all other rows, resulting in matrix $\boldsymbol{S}^b$ in Eq. (1.13). In the general case, this procedure would be continued until the first P columns of $\boldsymbol{S}$ have been processed, resulting in a matrix $\boldsymbol{S}^P$. Small non-zero values, marked by ε, are created in the process. They are ignored in this analysis (if desired, these values could be zeroed out by using the Eqs. (1.3) and (1.4)). A detailed analysis shows that taken together, all the rotations needed for obtaining $\boldsymbol{S}^P$ insert in the pth column of $\boldsymbol{U}$ a noise level corresponding to

$$\mathrm{S/N} = \sqrt{\frac{s_{pp}^2}{\sum_{i\neq p} s_{ip}^2}} \approx \sqrt{\frac{s_{pp}^2}{m-1}} \approx \sqrt{\frac{\sigma_{pp}^2}{m-1}} \tag{1.14}$$

Here, the notations s_{ip} and s_{pp} refer to elements of $\boldsymbol{S}$ in Eq. (1.8).

Eq. (1.14) demonstrates that the S/N ratio will be poorest for the last (Pth) principal component that has the smallest singular value, σ_{PP}. The success or failure of the PCA is determined by this singular value. For this reason, further discussion is limited to the last principal component only. Of course, this result is not new. It is clear to anybody doing PCA that the noise level of principal components increases with decreasing singular values.

2.2.3. The rotations for introducing zeroes in the first rows of S

The current representation of $\boldsymbol{X}$ is summarized as

$$\begin{aligned} \boldsymbol{X} &\approx \boldsymbol{U}^0\boldsymbol{U}^K\boldsymbol{S}^P\boldsymbol{V}^{0\mathrm{T}} \\ &= \boldsymbol{U}^0\boldsymbol{U}^K \begin{vmatrix} s_{11}^P & 0 & * & * & * \\ 0 & a & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{vmatrix} \boldsymbol{V}^{0\mathrm{T}} \end{aligned} \tag{1.15}$$

where the element s_{22}^P has been denoted by a. This element is approximately $a \approx \sqrt{\sigma_{22}^2 + m}$ (this follows from the fact that rotations between rows conserve the sum of squares of elements within each column).

Next, the upper right block of $\boldsymbol{S}^P$ should be zeroed. This process cannot be done by straightforward column rotations because non-zero elements of significant size would then be reintroduced in the leftmost columns. As preparation for efficient zeroing of the upper right elements, the lower right block is first rotated to a diagonal shape. This step is the key to success in the present analysis.

The matrix $\boldsymbol{S}^P$ of Eq. (1.15) is transformed as follows:

$$\boldsymbol{S}^P = \begin{vmatrix} s^P_{11} & 0 & \tilde{*} \\ 0 & a & \tilde{*} \\ \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{E}} \end{vmatrix} = \begin{vmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{U}} \end{vmatrix} \begin{vmatrix} s^P_{11} & 0 & \tilde{*} \\ 0 & a & \tilde{*} \\ \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{S}} \end{vmatrix} \times \begin{vmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{V}}^{\mathrm{T}} \end{vmatrix} = \boldsymbol{U}^Q \boldsymbol{S}^Q \boldsymbol{V}^{Q\mathrm{T}} \tag{1.16}$$

where $\tilde{\boldsymbol{E}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{V}}^{\mathrm{T}}$ is the SVD of $\tilde{\boldsymbol{E}}$. The rotation matrices $\boldsymbol{U}^Q$ and $\boldsymbol{V}^Q$ are not of interest as they only transform the basis of the noise part. The matrix $\boldsymbol{S}^Q$ can be written as

$$\boldsymbol{S}^Q = \begin{vmatrix} s^Q_{11} & 0 & * & * & * \\ 0 & a & b & * & * \\ 0 & 0 & d & 0 & 0 \\ 0 & 0 & 0 & s^Q_{44} & 0 \\ 0 & 0 & 0 & 0 & s^Q_{55} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} \tag{1.17}$$

One of the random elements in the upper right block has been singled out by denoting it by b. Similarly, the largest diagonal element of the lower right block has been denoted by d. The element b, and the elements marked by $(*)$ still possess the statistical properties of the original error matrix $\boldsymbol{E}$, i.e. they are $N(\mathbf{0}, \boldsymbol{I})$. As will be shown in detail, the element b, connecting the smallest "signal" singular value a with the largest "noise" singular value d, is the most critical of all remaining random values.

It might happen that $d > a$ in $\boldsymbol{S}^Q$. In such a case, obtaining the SVD would next require rotations that interchange the columns P and $P+1$, and similarly the rows P and $P+1$, of $\boldsymbol{S}^Q$. In such a case there would be more noise than signal in the last principal component. Thus, it is seen that PCA fails if there is so much noise that $d > a$. In the following, it is assumed that $d < a$ in $\boldsymbol{S}^Q$.

The following step needs to zero out the element marked by b by diagonalizing the 2 by 2 matrix block

$$\begin{vmatrix} a & b \\ 0 & d \end{vmatrix}$$

To accomplish this task requires both column and row rotations, i.e. rotations from right *and* from left, as shown by Eqs. (1.3) and (1.4). Taking into account that $f = 0$, the parameters p and q for these rotations are obtained from

$$p = \frac{a^2 - d^2 + b^2}{bd}, \qquad q = \frac{a^2 - d^2 - b^2}{ab} \tag{1.18}$$

The dependence of the rotational coefficient s on parameter q, as given by equation $s^2 + qs - 1 = 0$, is illustrated by the following equation:

q (or p)	1	2	3	5	10	100
s (or t)	0.6	0.4	0.3	0.19	0.1	0.01
S/N	1.6	2.4	3.3	5.2	10	100

(1.19)

It is seen that S/N decreases almost proportionally to q as q decreases towards zero (without loss of generality, it is assumed that $b > 0$, in order to simplify the presentation). It is easily seen that $q < p$. Thus, q controls the S/N of the $\boldsymbol{V}$ or right-hand factor elements and is the critical parameter. The following discussion concentrates on q. Typical values might be $a = 10$, $b = 1$, and $d = 5, 6, 7, 8, 9, 9.5$. The following table shows the q and S/N for these values.

d	5	6	7	8	9	9.5
q	7.4	6.3	5	3.5	1.8	0.9
S/N	7.5	6.4	5.2	3.8	2.2	1.5

(1.20)

The value of S/N drops quickly when d approaches a. With $d = 0.8a$, the signal still dominates clearly and the results would probably be considered useful. However, if the largest noise singular value $d = 0.95a$, then the signal is barely above noise and the result would hardly be useful any more. The essential result is that the critical range of d is narrow, on the order of the top 20%. When d grows over such a range, the results turn from good to bad.

The preceding discussion was about zeroing one single element from the upper right block. The effect of zeroing the other elements can be evaluated in a similar way. Their effect is usually less critical because the other differences between the signal singular value and the noise singular value will be larger than the difference between a and d above.

2.2.4. Scaling of variables (columns) or of observations (rows)

The analytic results were based on the assumption that $\mathrm{var}(e_{ij}) = 1$ for all i and j, i.e. that all errors are drawn from a $N(\mathbf{0}, \boldsymbol{I})$ distribution. Often one wants to analyze what happens if this assumption is not met: perhaps one has made a wrong assumption about the error variances, or intentionally scaled the variables differently, or not considered the error variances at all. The last alternative is true if so-called autoscaling is used, as will be explained later. The previous analysis is not quantitatively valid if the assumptions are not met. However, there is no specific reason to doubt the qualitative or approximate validity of the results. The results will be applied as a guideline for formulating numerical experiments that do not depend any further on the assumptions. Such numerical results will corroborate the qualitative validity of the analytic results in situations where the assumptions are not fully met.

The matrix to be scaled is $\boldsymbol{X} = \boldsymbol{X}^0 + \boldsymbol{E}^0$, i.e. the sum of an error-free data matrix and an error matrix consisting of $N(\mathbf{0}, \boldsymbol{I})$ random values. The equation for *column scaling* is

$$\boldsymbol{X}^d = \boldsymbol{X}\,\mathrm{diag}(\boldsymbol{d}) = \boldsymbol{X}^0\,\mathrm{diag}(\boldsymbol{d}) + \boldsymbol{E}^0\,\mathrm{diag}(\boldsymbol{d}) \quad (1.21)$$

or in element form

$$x^d_{ij} = d_j x_{ij} = d_j x^0_{ij} + d_j e^0_{ij} \quad (1.22)$$

The elements, d_j, of the scaling vector $\boldsymbol{d}$ act as scaling coefficients for columns $\boldsymbol{x}_j$ of the matrix. For column variances, a similar equation is approximately valid:

$$\mathrm{var}(\boldsymbol{x}^d_j) = d_j^2\,\mathrm{var}(\boldsymbol{x}_j) \approx d_j^2\,\mathrm{var}(\boldsymbol{x}^0_j) + d_j^2\,\mathrm{var}(\boldsymbol{e}^0_j) \quad (1.23)$$

where the subscripted vectors denote columns of the corresponding matrices. The motivation of scaling is to minimize the effect of noise without unduly suppressing true information. In real-life situations, $\boldsymbol{d}$ must be chosen without detailed knowledge of the values x^0_{ij} and e^0_{ij}. Different approaches are possible for choosing $\boldsymbol{d}$. In the customary technique of *autoscaling*, the values d_j are chosen so that for each column j, $\mathrm{var}(\boldsymbol{x}^d_j) = 1$. It is easily seen from Eq. (1.23) that then the variances of the columns of the scaled error matrix $\boldsymbol{E}^0\,\mathrm{diag}(\boldsymbol{d})$ can be very different. Consider two different variables (columns); one with S/N = 1 (i.e. small $\mathrm{var}(\boldsymbol{x}^0_j)$) and another with S/N = 10 (i.e. with 10% error level, large $\mathrm{var}(\boldsymbol{x}^0_j)$). For the first one, the variance of the error part in the scaled matrix $\boldsymbol{X}^d$ is 0.5. For the second one, the variance is ≈0.01. The elements of the scaled error matrix are very far from being $N(\mathbf{0}, \boldsymbol{I})$. Subsequently, it will be demonstrated that this scaling prevents seeing weak factors because the noise from the low-S/N columns masks the weak factors.

It has also been suggested that the values d_j be chosen so that the variances of the columns of the scaled error matrix $d_j e^0_{ij}$ be equal [5]. Such scaling requires that some information be available about the sizes of errors of different variables.

The purpose of the present work is to explore the consequences of different scaling approaches and to suggest principles of column scaling, to be applied in real-life analyses.

Row scaling is also possible. Sometimes sample compositions are expressed as percentages of sampled mass. This represents one example of row scaling: the sum of concentrations in each row of the data matrix is scaled to unity. The equation for row scaling is

$$x^d_{ij} = d_i x_{ij} = d_i x^0_{ij} + d_i e^0_{ij} \quad (1.24)$$

Row scaling is only discussed in passing in the present work. Application of row scaling is problematic when environmental concentrations are analyzed. Samples of low and high concentration may represent different environmental conditions and weighting them may bias the results. Without understanding the questions connected with unwanted biasing, row weighting is not a safe approach when environmental concentrations are analyzed.

2.3. Influence of scaling on singular values of true data

Scaling a column will influence both the errors and the data contained in the values.

In the SVD of column-scaled true data, defined by $\boldsymbol{X}^0\,\mathrm{diag}(\boldsymbol{d}) = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$, the dependence of each "data" singular value s_{pp} on d_j is given by

$$\frac{\partial(s_{pp})}{\partial d_j} \approx s_{pp} v^2_{jp} \quad (1.25)$$

in the neighborhood of $d_j = 1$. The "*loadings*", i.e. the values v_{jp} indicate how much each variable j

participates in formulating factor p. Eq. (1.25) indicates that scaling variables with small loadings v_{jp} down has a very slight effect on singular value p. Such variables can be scaled down without adverse effects on the singular values corresponding to true data.

An equation analogous to (1.25) can be formulated for describing how scaling of matrix rows influences the singular values. It is seen that scaling the rows downward with small scores u_{ip} does not have much effect on the pth singular value.

3. Singular values of noise matrices

3.1. Matrices containing $N(\mathbf{0}, \boldsymbol{I})$ pseudorandom values

A numerical study of the largest singular values was conducted as follows. The largest singular value s_{11} was computed for a number of matrices $\boldsymbol{E}$ of dimensions, $m \times n$, containing $N(\mathbf{0}, \boldsymbol{I})$ values. For this study, the ranges of dimensions m and n were chosen as $40 \leq m \leq 100$, $10 \leq n \leq 30$. It was found that within this range, the average value of s_{11} is well approximated by the expression

$$s_{11} \approx \sqrt{m} + \sqrt{n} - 0.5 \tag{1.26}$$

The bias of the approximation (1.26) is less than 0.3, while the half-width of the distribution of actual values of s_{11} (for any given values m and n) is on the order of 0.7. Thus, this approximation is good enough for the purposes of the present work. The result of using Eq. (1.26) gives an estimate of the quantity d in Eqs. (1.17) and (1.18) for situations where the errors of data are $\boldsymbol{N}(\mathbf{0}, \boldsymbol{I})$. Considering that the dimension of $\tilde{\boldsymbol{E}}$ in (1.16) is $(m - P) \times (n - P)$, the estimate is $d \approx \sqrt{m - P} + \sqrt{n - P} - 0.5$, where m and n are the dimensions of the data matrix $\boldsymbol{X}$.

3.2. Matrices containing different amounts of noise in different columns

In practice, different columns of $\boldsymbol{X}$ often contain different amounts of noise. In order to study this situation, the largest singular value s_{11} was computed for a number of column-scaled matrices $\boldsymbol{E} = \boldsymbol{E}^0 \,\mathrm{diag}(\boldsymbol{d})$, where the elements of $\boldsymbol{E}^0$ are assumed to be $N(\mathbf{0}, \boldsymbol{I})$. The elements of vector $\boldsymbol{d}$ act as scale coefficients setting the standard deviations of columns of $\boldsymbol{E}$. The dimensions of $\boldsymbol{E}$ were chosen to be 100×20.

The first test was performed so that the first element d_1 of $\boldsymbol{d}$ was varied from 0 to 1.6, while the other elements of $\boldsymbol{d}$ were kept at unity. For a number of randomly generated test matrices, the behavior of s_{11} was recorded as a function of d_1. The results are shown in Fig. 1. Considerable variation from one test matrix to another is visible. On the average, s_{11} decreases very slightly when the scale coefficient d_1 decreases from 1 to 0. In contrast, increase of d_1 above unity causes a dramatic increase of s_{11}. The onset of this increase varies from case to case.

The second test was performed so that the scale coefficients d_1 to d_5 were varied together. Now there is less case-to-case variation (see Fig. 2). Otherwise, the results are similar: decreasing the scale of five columns from 1 to 0 decreases s11 by approximately 0.5 units. In contrast, an increase of the five scale coefficients from 1 to 1.6 increases s_{11} by 4 or 5 units.

These results lead to formulating practical advice. As a starting point in the assumed practical situation, it is assumed that information about errors of different variables has already been utilized for scaling the

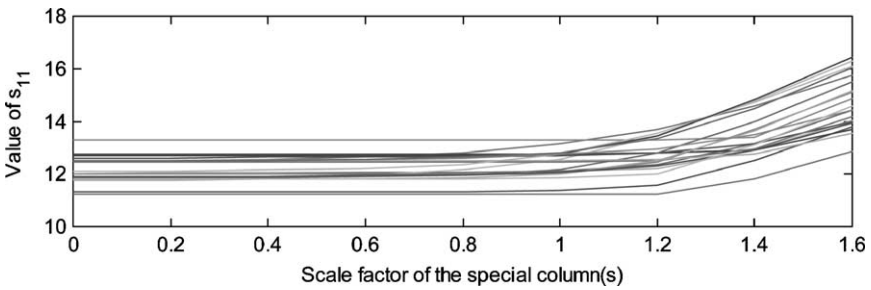

Fig. 1. s_{11} of noise as a function of the scale of one column (scale factor of the special column(s)).

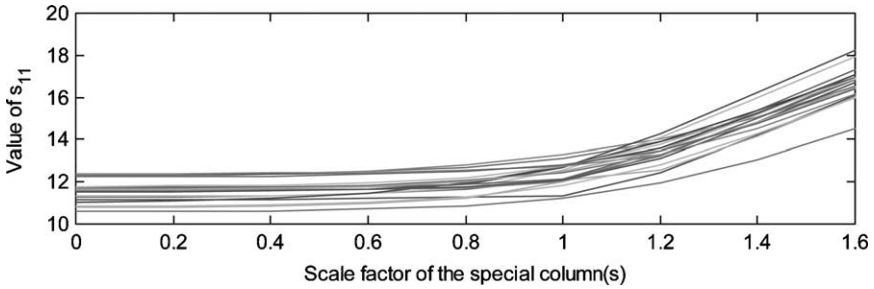

Fig. 2. s_{11} of noise as a function of the scale of five columns (scale factor of the special column(s)).

variables so that variances of errors in different columns of $\boldsymbol{X}$ are believed to be unity. If the estimation of errors has been reliable, then this belief is essentially true. The largest singular value of the error matrix can then be slightly decreased by scaling down some variables by a coefficient of 0.5, say. The variables with small values v_{jP} are possible candidates for downward scaling. Eq. (1.25) indicates that the Pth data singular value will not decrease much if such variables are scaled down. Thus, some gain may be achieved, i.e. the difference between data singular values and noise singular values may be increased. However, achieving this gain depends on the random details of the matrices. In some cases, no gain is achieved, as shown by Figs. 1 and 2.

In many practical situations, however, the estimation of errors is not reliable, especially for the weakest variables. Then it may be that in the initial data processing, some variables have in fact been scaled up, so that their error variances are significantly above unity. Subsequent downward scaling of such variables will significantly decrease the largest singular value of noise. For example, if the errors in the five variables had been underestimated by a factor of 1.6, then scaling them down by a factor of 0.5 will decrease the largest noise singular value by 5 units in the cases shown in Fig. 2. It appears that as a safeguard against uncertainties in error estimation, all weak variables should be scaled down by a factor of 0.5.

The results of this section demonstrate that the largest singular value of noise is strongly influenced by the columns that have the largest errors. It is seen that autoscaling behaves badly: it inflates the singular values of noise by scaling up the errors in the variables with low S/N. The inflated noise may entirely mask the weakest factor(s).

4. Numerical experiments with matrices containing signal and noise

In this section, a realistic experimental plan is worked through. First, the data matrix $\boldsymbol{X}^0$ and an error matrix $\boldsymbol{E}$ are set up. The elements of $\boldsymbol{E}$ are generated to be $N(\boldsymbol{0}, \boldsymbol{I})$. The PCA of $\boldsymbol{X}^0$ is computed as $\boldsymbol{X}^0 = \boldsymbol{U}^0\boldsymbol{S}^0\boldsymbol{V}^{0\mathrm{T}}$. The matrices $\boldsymbol{U}^0$ and $\boldsymbol{V}^0$ are set aside as a reference. Also, the PCA of $\boldsymbol{X} = \boldsymbol{X}^0 + \boldsymbol{E}$ is computed as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$. The differences between $\boldsymbol{U}^0$ and $\boldsymbol{U}$, and between $\boldsymbol{V}^0$ and $\boldsymbol{V}$, represent the effect of noise on the scores and loadings. These differences specify the "benchmark" level, the level of performance that is obtained when the "standard" approach is taken. Finally, the SVD of several different column-scaled matrices is computed as $\boldsymbol{X}^d = \boldsymbol{U}^d\boldsymbol{S}^d\boldsymbol{V}^{d\mathrm{T}}$. The effect of different scaling alternatives is then indicated by the differences between $\boldsymbol{U}^0$ and $\boldsymbol{U}^d$, and between $\boldsymbol{V}^0$ and $\boldsymbol{V}^d$.

4.1. Computing a measure of difference between two versions of a factor matrix

In this section, the notation is simplified so that $\boldsymbol{U}$ and $\boldsymbol{V}$ denote the partial matrices obtained by only including the P first columns of the full $\boldsymbol{U}$ and $\boldsymbol{V}$.

The simple measure $\|\boldsymbol{U} - \boldsymbol{U}^0\|_F$ is sensitive to rotations within the P-dimensional subspace spanned by factors. In order to eliminate such sensitivity, the measure of distance between $\boldsymbol{U}^0$ and $\boldsymbol{U}$ is obtained as the minimum of $\|\boldsymbol{U} - \boldsymbol{U}^0\boldsymbol{Q}\|_F$ under the constraint that $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{I}$. This minimum is denoted by $|\boldsymbol{U} - \boldsymbol{U}^0|_{\mathrm{Proc}}$ in the following. The minimization problem is called *the orthogonal Procrustes problem* [4]. Procrustes rotations have also been used as a tool for solving factor analytic problems (e.g. [6]). The present application of Procrustes rotations is simpler. It is just a method

Table 1
Differences between the computed factor matrices ($\boldsymbol{U}$ and $\boldsymbol{V}$) and the true factor matrices ($\boldsymbol{U}^0$ and $\boldsymbol{V}^0$)

Scaling of five weak columns	$\min\|\boldsymbol{U}-\boldsymbol{U}^0\|_{\text{Proc}}$	$\max\|\boldsymbol{U}-\boldsymbol{U}^0\|_{\text{Proc}}$	$\min\|\boldsymbol{V}-\boldsymbol{V}^0\|_{\text{Proc}}$	$\max\|\boldsymbol{V}-\boldsymbol{V}^0\|_{\text{Proc}}$
Standard	0.49	0.73	0.19	0.40
Autoscaling	1.7	1.9	1.2	1.8
Down by 0.5	0.49	0.73	0.19	0.40
Up by 2.0	0.60	1.4	0.23	1.1

The notations min and max refer to smallest and largest values obtained in 20 randomized trials.

of measuring the distance of an approximate solution from the true solution.

The up- or down-scaling of columns of $\boldsymbol{X}$ has a direct influence on the corresponding rows of $\boldsymbol{V}$. In order to avoid complications, rows of $\boldsymbol{V}$ corresponding to weak columns of $\boldsymbol{X}$ were excluded from the comparisons between differently scaled versions of the $\boldsymbol{X}$ matrix.

4.2. Generation of test matrices

The dimensions of $\boldsymbol{X}$ were set to $m = 100$, $n = 30$. The first ten columns of $\boldsymbol{X}$ are set to represent *weak* variables with low S/N. The number of factors was set to $P = 3$. The elements of the true factor matrix $\boldsymbol{G}$ were generated as pseudorandom numbers, uniformly distributed between 0 and 2 (average = 1). The elements of the error matrix E were generated as normally distributed pseudorandom values with mean = 0, S.D. = 1, i.e. $N(\boldsymbol{0}, \boldsymbol{I})$.

Columns 11–30 of the true factor matrix $\boldsymbol{F}$ (the main part of $\boldsymbol{F}$) were filled with pseudorandom numbers, uniformly distributed between 0 and 3.33. This procedure causes the average size of elements in the main part of $\boldsymbol{X}$ to be 5. Thus for the main part of $\boldsymbol{X}$, S/N = 5. For the first ten columns of $\boldsymbol{F}$, the pseudorandom true values were uniformly distributed between 0 and 0.33. Thus, for the weak part of $\boldsymbol{X}$, S/N = 0.5, the test was replicated 20 times. For each test, all the random values ($\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{E}$) were generated anew with different pseudorandom seed values.

4.3. Results of the numerical test

The results are shown in Table 1. The min and max difference values relate to the smallest and largest differences found in 20 experiments performed with different random numbers. Difference values less than 1 indicate approximately the relative error in the elements of the weakest or *P*th factor vector. Difference values well above unity indicate that there is also a significant amount of error in the other factor vectors, and not only in the weakest factor.

The table indicates that $\boldsymbol{V}$ matrix elements are obtained with less error than the $\boldsymbol{U}$ elements. There is considerable spread from the smallest to the largest difference values. The main reason for this spread is that the condition numbers of true factor matrices vary randomly from case to case. The error level in the test was originally chosen so that the standard scaling gives marginally useful results, S/N for the weakest factor ranged from 1.5 to 5. The most dramatic result in Table 1 is that in all cases, autoscaling leads to useless results for both $\boldsymbol{U}$ and $\boldsymbol{V}$.

When scaling the five weak columns down by 0.5, the results changed only in the third digit when compared to standard scaling. Figs. 1 and 2 predict such a result. With more or less signal in the weak columns, the difference values for “down by 0.5” will probably deviate more from those of standard scaling: the difference values will increase with more signal, and decrease with less signal. In the ultimate case of no signal in the weak columns, scaling down is of course expected to decrease the difference values.

The opposite scaling, scaling up the five weak columns by 2.0, produces large errors both in $\boldsymbol{U}$ and in $\boldsymbol{V}$, again as predicted by Figs. 1 and 2. In some cases, the weakest $\boldsymbol{U}$ and $\boldsymbol{V}$ factors lose all observable signal.

5. Discussion

In publications where PCA is applied, one may occasionally observe the belief that there should be no signal in the rejected singular components, provided

that the number of significant singular components has been correctly determined. The analysis based on Givens rotations clearly demonstrates that this belief is false. The rotations leading up to the SVD of the noise-containing matrix will mix signal and noise. Noise will get rotated into the significant components and simultaneously signal will get rotated into the components that will be rejected.

There is no way to recover the entire signal from the components to be rejected. However, with a better scaling of weak and/or bad components, the mixing of noise into the signal can be minimized so that a minimum of signal is wasted and simultaneously a minimum of noise will contaminate the significant components.

5.1. Do not autoscale noisy variables in PCA

In PCA, it is customary to scale columns of $\boldsymbol{X}$ so that in the scaled matrix, all of the columns have the same variance. This procedure means that the sum of the two components of the variance (signal and noise) is constant over all variables. It follows that for the weakest variables, having the smallest amount of signal, the noise variance is much larger than for the strong variables. This behavior is in severe conflict with the recommendations found in this work: the exaggerated noise in a few noisy variables will cause the small principal components to be undetectable in the analysis and will increase the noise in other principal components. The recommendation is clear: "do not auto-scale noisy variables in PCA modeling". Instead, it would be best to scale variables so that the noise variance is the same in all variables. In order to be on the safe side, it is advisable to scale the weak variables so that their noise variance is on the order of 30–50% of the noise variance of the stronger variables. In environmental work, one should probably scale many trace element concentrations in this way.

By default, commercial statistical packages apply autoscaling as a preparatory step of PCA. Usually an option is available for turning autoscaling off although the option may not be easy to find. If autoscaling is omitted, then the user has to scale the data properly before applying PCA and "unscale" the results computed by PCA. In practice, autoscaling may sometimes be the only practical option when using PCA contained in a software package. In such a situation, it is definitely best to discard such variables that do not display a clear signal. It is not possible to state a precise limit for discarding variables in such a case. Typically, one should discard variables where noise variance exceeds signal variance (i.e. $S/N < 1$). On the other hand, it is unlikely that variables with $S/N > 3$ should be discarded. Between these limits, experiments might be performed with different schemes of rejecting variables. For censored variables, the rejection of noisy variables might be based on the criterion for weak variables, as presented in Eq. (1.1).

If autoscaling is turned off but available resources do not allow the determination of noise levels, then the following rule of thumb can be applied in environmental PCA analysis. Let $v_j = \sqrt{\mathrm{var}(\boldsymbol{x}_j)}$ be the standard deviation of column j of the original data matrix. For strong variables, compute the scaled matrix elements x_{ij}^d from the equation $x_{ij}^d = x_{ij}/v_j$. In this way the main (strong) variables are scaled to unit variance. For variables where noise is expected to be of similar magnitude as signal, compute x_{ij}^d from the equation $x_{ij}^d = x_{ij}/4v_j$.

Finally, for those variables where hardly any signal is present, either omit such variables from the analysis, or use $x_{ij}^d = x_{ij}/20v_j$.

5.2. Recommendations for positive matrix factorization

Column scaling of variables is not used with PMF. Instead, an uncertainty is specified for each individual data value. If these uncertainties are specified too small or too large in comparison to the true error level of a certain variable, then an over- or downweighting of the variable occurs. Regarding weak/bad variables, the main result of this work is that even a small amount of overweighting is quite harmful and should be avoided. In contrast, moderate downweighting, by a factor of 2 or 3, never hurts much and sometimes is useful. Thus, it is recommended to routinely downweight all weak variables by a factor of 2 or 3. This practice will act as insurance and protect against occasions when the error level of some variables has been underestimated resulting in a risk of overweighting such variables. Regarding bad variables (where hardly any signal is visible from the noise), the recommendation is that such variables be entirely omitted from the model. If this is not desirable, then such variables

should be strongly downweighted, by a factor of 5 or 10.

5.3. Validity of these results in analysis of environmental data

The detailed results obtained in this work were based on the following assumptions: (1) errors are statistically independent; (2) errors are normally distributed with expected value = 0; (3) within each single column, the data are homoscedastic, i.e. data errors have the same standard deviation for each value in the column. None of these assumptions is expected to hold for environmental measurements. The assumption of independence is probably the most crucial one. Factor analytic methods analyze features common to several variables. Hence, errors common to several variables are more likely to influence the results than independent errors of the same magnitude. Thus, it is possible or even likely that in reality the need to downweight or reject weak/bad variables is more acute than predicted by the detailed results found in this work. This question can be studied by simulation studies to some extent. Unfortunately, the true nature of errors is not well known and simulation would lack a solid basis for choosing an error structure. An example of correlated errors may be found when the elemental composition of aerosol samples is determined by multivariate methods such as X-ray fluorescence (XRF). Subtracting the background under the peaks is critical for the determination of lighter elements, such as Na, Mg, Al, Si and P. Errors in background fitting will cause correlated errors in these adjacent elements. Downweighting of these elements might be necessary even although their concentrations might be significant in a large majority of all samples. Gaarenstroom et al. [7] observed correlated errors in instrumental neutron activation analysis and similar problems can be envisioned in other analytical systems.

It is hard to give a general estimate of the significance of the assumptions (2) and (3) above. In any practical situation, these questions can be studied by simulation. Practical experience with environmental data suggests that dramatic results are not to be expected from such simulations.

Sometimes an important variable of an experiment may appear noisy. Then it is important to know whether the variable can be retained in the analysis. In such a case, special care is needed for producing reliable error estimates for such data. It is strongly advised that parallel sampling or parallel analyses of a fraction of all data be performed in order to verify error levels reported by the analyst. Parallel sampling is also useful in order to monitor variability connected with sampling.

5.4. Planning environmental measurements

The authors have encountered aerosol data sets where more than half of all elemental concentrations had to be classified as weak or bad. The cost of measuring such concentrations is essentially wasted. It is suggested that major studies should be preceded by a preliminary phase where the S/N ratio of concentrations is surveyed. Elements with a low S/N should be excluded from the study if savings in the cost of chemical analyses can be realized in this way. The tradeoff of performing parallel sampling as compared with measuring additional weak variables should be carefully considered. It is likely that the increased reliability achieved by parallel sampling and/or by parallel analyses will outweigh the additional information that might possibly be gained by measuring additional but weak variables.

Acknowledgements

This work was supported in part by Unilever Research US. The authors would like to thank Prof. William Rayens of the University of Kentucky for useful discussions on this subject.

Appendix A. Givens rotations

Givens rotations are one of the main tools in matrix computations [4]. The general form of a Givens rotation is as follows:

$$\boldsymbol{H} = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & s & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -s & 0 & 0 & c & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} \tag{A.1.27}$$

where $c^2 + s^2 = 1$. Givens rotations are orthonormal: $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{I}$. Givens rotations can be applied for modifying two matrices such that the product of the two matrices does not change, in the following way: $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})\boldsymbol{B} = (\boldsymbol{A}\boldsymbol{H}^{\mathrm{T}})\boldsymbol{H}\boldsymbol{B} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}$. In this example, matrix $\boldsymbol{B}$ is rotated from left by Givens rotation $\boldsymbol{H}$. In order that the product $\boldsymbol{A}\boldsymbol{B}$ does not change, matrix $\boldsymbol{A}$ must be rotated from the right by the transpose of $\boldsymbol{H}$. By performing the multiplication, one can see that the matrices $\boldsymbol{B}$ and $\tilde{\boldsymbol{B}}$ are identical except for two rows. Similarly, the matrices $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ are identical except for two columns.

Givens rotations are used for transforming a matrix into simpler structure. Often, each rotation creates a zero-valued element in the rotated matrix. Terminology: it may be said that the example rotation in Eq. (A.1.27) rotates column 2 with column 5, or rotates columns 2 and 5, or rotates column 5 into column 2. The last form might be used if the emphasis (the element to be zeroed) is in column 2.

The diagonal elements of the product matrix $\mathbf{A}^{\mathrm{T}}\boldsymbol{A}$ equal the sums of squares of elements over columns of $\boldsymbol{A}$. Application of a Givens rotation from the left gives $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{A} = \tilde{\boldsymbol{A}}^{\mathrm{T}}\tilde{\boldsymbol{A}}$. It follows that a Givens rotation from the left does not change the sum of squares of elements in any column of the matrix. Similarly, a Givens rotation from the right conserves the sums of squares of elements on the rows of the matrix.

References

[1] S. Huang, K. Rahn, R. Arimoto, Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island, Atmos. Environ. 33 (1999) 2169–2185.
[2] 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.
[3] P. Paatero, Least squares formulation of robust non-negative factor analysis, Chemometrics Intelligent Lab. Syst. 37 (1997) 23–35.
[4] G.H. Golub, C.F. Van Loan, Matrix Computations, First ed., North Oxford Academic, Oxford, 1983.
[5] P. Paatero, U. Tapper, Analysis of different modes of factor analysis as least squares fit problems, Chemometrics Intelligent Lab. Syst. 18 (1993) 183–194.
[6] X. Tomas, J.M. Andrade, A. Alvarez-Larena, Chemometric analysis of skeletal data from non-fused and non-pi-complexed pentafulvenes, Talanta 48 (1999) 781–794.
[7] P.D. Gaarenstroom, S.P. Perone, J.P. Moyers, Application of pattern recognition and factor analysis for characterization of atmospheric particulate composition in southwest desert atmosphere, Environ. Sci. Technol. 11 (1977) 795–800.