JOURNAL OF CHEMOMETRICS
J. Chemometrics 2003; **17**: 16–33
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.773

Centering and scaling in component analysis†

Rasmus Bro[1]* and Age K. Smilde[2]

[1]Chemometrics Group, Food Technology, Department of Dairy and Food Science, Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark
[2]Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, NL-1018 WV Amsterdam, The Netherlands

Received 10 May 1999; Accepted 1 June 2001

In this paper the purpose and use of centering and scaling are discussed in depth. The main focus is on two-way bilinear data analysis, but the results can easily be generalized to multiway data analysis. In fact, one of the scopes of this paper is to show that if two-way centering and scaling are understood, then multiway centering and scaling is quite straightforward. In the literature it is often stated that preprocessing of multiway arrays is difficult, but here it is shown that most of the difficulties do not pertain to three- and higher-way modeling in particular. It is shown that centering is most conveniently seen as a projection step, where the data are projected onto certain well-defined spaces within a given mode. This view of centering helps to explain why, for example, centering data with missing elements is likely to be suboptimal if there are many missing elements. Building a model for data consists of two parts: postulating a structural model and using a method to estimate the parameters. Centering has to do with the first part: when centering, a model including offsets is postulated. Scaling has to do with the second part: when scaling, another way of fitting the model is employed. It is shown that centering is simply a convenient technique to estimate model parameters for models with certain offsets, but this does not work for all types of offsets. It is also shown that scaling is a way to fit models with a weighted least squares loss function and that sometimes this change in objective function cannot be performed by a simple scaling step. Further practical aspects of and alternatives to centering and scaling are discussed, and examples are used throughout to show that the conclusions in the paper are not only of theoretical interest but can have an impact on practical data analysis. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: offset; weighted least squares; preprocessing; two-way; three-way; multiway; missing data; PCA; PARAFAC

1. INTRODUCTION

1.1. Definitions

It is important to have a concise terminology for scaling and centering. The following convention is based on suggestions from the literature [1–4]. The term 'an offset'—also sometimes called an intercept—is used for a part of the model that is constant across one or several modes. An R-component bilinear model of a data matrix $\mathbf{X}$ ($I \times J$) with elements x_{ij} may be written in terms of scalars or in matrix notation as

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{E} \Leftrightarrow x_{ij} = \sum_{r=1}^{R} \phi_{ir}\theta_{jr} + \varepsilon_{ij} \quad (1)$$

where Φ ($I \times R$) and Θ ($J \times R$) hold the parameters ϕ_{ir} and θ_{jr} respectively and Greek letters are used to indicate population parameters. The matrix $\mathbf{E}$ holds the unknown errors. Offsets may be constant across the first mode (rows). The model associated with such offsets is

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{E} \Leftrightarrow x_{ij} = \sum_{r=1}^{R} \phi_{ir}\theta_{jr} + \mu_j + \varepsilon_{ij} \quad (2)$$

where $\boldsymbol{\mu}$ ($J \times 1$) holds the constant terms μ_j ($j = 1,\ldots,J$), and $\mathbf{1}$ is a one-vector of suitable size ($I \times 1$ in this case). Again the Greek letter $\boldsymbol{\mu}$ indicates a population value. Offsets may

*Correspondence to: R. Bro, Chemometrics Group, Food Technology, Department of Dairy and Food Science, Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark.
E-mail: Rasmus.bro@optimax.dk
†Dedicated to Professor John F. MacGregor: a pioneer of multivariate statistical process control and recipient of the fourth Herman Wold medal.
Contract/grant sponsor: LMC (Center for Advanced Food Studies).
Contract/grant sponsor: EU (European Union); Contract/grant number: NwayQual GRD1-1999-10377.
Contract/grant sponsor: AQM (Advanced Quality Monitoring)/Danish Ministries of Research and Industry.

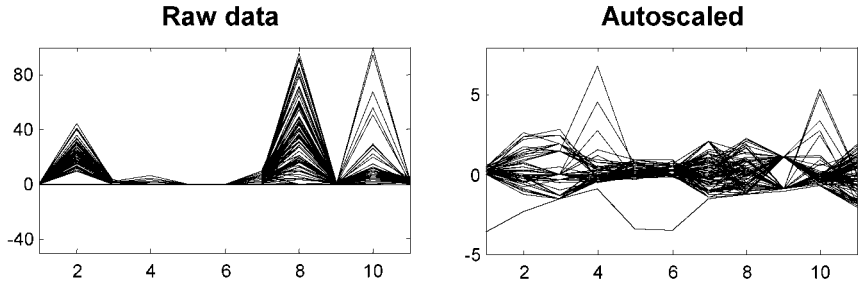

Figure 1. Example of centering and scaling using sugar process data (see text).

also be constant across the second mode (columns). The underlying bilinear model with offsets across the second mode reads

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \boldsymbol{\mu}\mathbf{1}^{\mathrm{T}} + \mathbf{E} \Leftrightarrow x_{ij} = \sum_{r=1}^{R} \phi_{ir}\theta_{jr} + \mu_i + \varepsilon_{ij} \quad (3)$$

where the vector $\boldsymbol{\mu}$ ($I \times 1$) is now holding the offsets μ_i ($i = 1,\ldots,I$). Offsets may also be constant across columns *and* across rows, yielding

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mu\mathbf{1}\mathbf{1}^{\mathrm{T}} + \mathbf{E} \Leftrightarrow x_{ij} = \sum_{r=1}^{R} \phi_{ir}\theta_{jr} + \mu + \varepsilon_{ij} \quad (4)$$

in which the single constant μ is the same for all elements of $\mathbf{X}$. Such a situation may arise for example in chromatography or capillary electrophoresis, where a constant offset in the detector may appear owing to the way in which the detector zero-level is determined.

Thus for bilinear models there are two basic types of offsets: constants across one mode (columns or rows) or constants across both modes. Combinations of such offsets may also appear, as seen for example in analysis-of-variance settings.

As will be shown below, offsets are often handled by first centering the data and subsequently fitting the bilinear model. If the data are centered by subtracting the column average from every element in the column, this is referred to as *centering across the first mode*. Mathematically it can be expressed as

$$y_{ij} = x_{ij} - \frac{\sum_{i=1}^{I} x_{ij}}{I} \quad (5)$$

where y_{ij} is an element of the centered data matrix. If $\mathbf{m}$ ($J \times 1$) is a vector holding the jth column average in its jth element, then centering across the first mode can also be expressed as

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}\mathbf{m}^{\mathrm{T}} \quad (6)$$

where $\mathbf{1}$ is an I-vector of ones and $\mathbf{Y}$ is the matrix holding the centered data. Subtracting the row average from each element in a row is referred to as *centering across the second mode* and can be expressed as

$$y_{ij} = x_{ij} - \frac{\sum_{j=1}^{J} x_{ij}}{J} \quad (7)$$

or, using $\mathbf{m}$ ($I \times 1$) as a vector holding the ith row average in its ith element,

$$\mathbf{Y} = \mathbf{X} - \mathbf{m}\mathbf{1}^{\mathrm{T}} \quad (8)$$

In general, centering across one mode is also called *single centering*, and performing for example a centering across the first mode and then a subsequent centering *of the outcome* across the second mode is called *double centering*. The term slab centering, which is sometimes seen in the literature, refers to centering by subtracting, from each slab in a three-way array, the overall average of that slab. For two-way data this simply corresponds to subtracting the average of all elements.

For scaling, another terminology is used. When a matrix is scaled such that each row is multiplied by a specific scalar, it is called *scaling within the first mode* ($y_{ij} = x_{ij}w_i$). If each column is multiplied by a certain scalar as in traditional autoscaling, it is referred to as *scaling within the second mode* ($y_{ij} = x_{ij}w_j$). In matrix notation, scaling within the first mode can be written as

$$\mathbf{Y} = \mathbf{W}\mathbf{X} \quad (9)$$

where $\mathbf{W}$ is an $I \times I$ diagonal matrix holding the scalar w_i in its ith diagonal element. Scaling within the second mode can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{W} \quad (10)$$

where $\mathbf{W}$ is now a $J \times J$ diagonal matrix holding the scalar w_j in its jth diagonal element.

An example of centering and scaling is shown in Figure 1. The data are from a sugar factory (see Reference [5] for more information). The variables shown are ash (1), color (2), color type (3), turbidity (4), grain sizel (5), grain size2 (6), SO_2 (7), invert (8), floc (9), insoluble residue (10) and amino-N (11), which are all measured in different units of different magnitude. Each line in a plot is the status ('spectrum') of these 11 variables at a certain time. Ninety-seven times are shown. The raw data are shown on the left side. The different ranges of these variables will manifest themselves in a subsequent modeling of the data, where the variables with little variation will not be modeled to any significant degree. Centering (across the first mode) will not remove these scale differences but will move the variation of each variable to the zero-level. As the differences in scales between variables are arbitrary, it is useful to scale the data so that each variable has the same initial standard deviation (and remove different measurement units). This can be achieved by scaling the centered data within the second (variable) mode.

The corresponding *autoscaled* data are shown on the right side of the figure. It is seen that the variation of each processed variable is comparable for the autoscaled data. The outlying sample in the lower part of the plot, however, leads to too dramatic a downweighting of, for example, variable 1 and should hence be excluded before preprocessing is carried out.

It may seem strange that different words are used for scaling (within) and centering (across). The explanation for this is as follows. Centering is performed across a mode in the sense that *one* offset is subtracted from every element in a certain vector, i.e. the data are centered *across* the elements of one mode. The same holds for three-way data; the average value is subtracted from each element of a vector. Scaling is performed by multiplying *all* elements in the array containing a certain variable (or object) by the same scalar. For two-way data, scaling therefore also pertains to vectors, but e.g. for three-way data this means that a whole slab (corresponding for example to the $I \times K$ matrix of the jth wavelength in a spectral three-way array) has to be multiplied by the same scalar. Thus scaling is performed *within the elements* of one mode.

In the following, much use will be made of the notion of 'fit' or 'model fit'. In general terms this means what portion of the data is fitted by the model. This can be expressed by

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E}$$
$$R^2 = \frac{\|\mathbf{X}\|^2 - \|\mathbf{E}\|^2}{\|\mathbf{X}\|^2} = 1 - \frac{\|\mathbf{E}\|^2}{\|\mathbf{X}\|^2} \tag{11}$$

where $\mathbf{X}$ contains the data, $\hat{\mathbf{X}}$ contains the fitted values of the data using the model, and $\mathbf{E}$ contains the residuals. The symbol $\|\mathbf{A}\|$ denotes a norm of $\mathbf{A}$, here taken to be the Frobenius or Euclidean norm (square root of the sum of squared elements of $\mathbf{A}$). The R^2 statistic can take on values between zero and one, where one means perfect fit and zero means no fit. The fit may also be expressed in percentages between 0% and 100%.

1.2. Leading Principles

There are two leading principles in this paper. The first principle is parsimony. It is preferred that a model is as simple as possible. This means that if two models give the same fit, the model using the fewer parameters is preferred. This idea goes back to William of Ockham who lived in the Middle Ages and stated that a minimum number of assumptions should be adopted to explain a phenomenon [6]. This principle is known as 'Ockham's razor'. In statistics the notion of parsimony is formulated in statistical decision theory [7], and in chemometrics it was introduced as the 'parsimony principle' [8].

The second principle is that centering in this paper is not considered to *estimate* offsets but to *remove* offsets. Estimating offsets is a different issue than removing them, and estimating offsets has its own properties and problems.

1.3. Outline of paper

In the main part of the paper, only two-way (bilinear) models are considered, because most results generalize straightforwardly to multiway models. Section 2 is concerned with two-way centering, Section 3 with two-way scaling, and Section 4 with the combined use of two-way centering and scaling. Section 5 explains the discussed results in terms of multiway models, and finally in Section 6 some conclusions are drawn.

1.4. Notation

The following notation is used in this paper. Two-way data are held in an $I \times J$ matrix $\mathbf{X}$ with typical elements x_{ij}. Three-way data are held in an $I \times J \times K$ matrix $\underline{\mathbf{X}}$ with typical elements x_{ijk}. Such a three-way array is often rearranged to an $I \times JK$ matrix $\mathbf{X}^{(I \times JK)}$ by concatenating the K third-mode frontal slabs after each other, i.e. $\mathbf{X}^{(I \times JK)} = [\mathbf{X}_1\ \mathbf{X}_2\ \ldots \mathbf{X}_K]$, where $\mathbf{X}_k$ is the $I \times J$ matrix obtained by fixing the third mode of $\underline{\mathbf{X}}$ at k. This operation has been referred to in a number of ways (e.g. unfolding), but it has been suggested to use the term matricizing to avoid confusion with other techniques [2]. The matricized array is often just denoted $\mathbf{X}$ instead of $\mathbf{X}^{(I \times JK)}$ if no confusion is possible. The letter $\mathbf{Y}$ is used for preprocessed data and $\hat{\mathbf{X}}$ is used for a model of $\mathbf{X}$ (be it two- or three-way). The number of components in a model is called R.

The letter m is used for calculated offsets (e.g. averages) and μ is used for true offsets (population values). The letter w is used for a scaling parameter. The letter $\mathbf{P}$ is used for a projection matrix related to centering. Usually its dimension will not be specified, because it follows from the context. Similar rules apply for the diagonal matrix $\mathbf{W}$ holding the weights associated with scaling the array.

The Kronecker product is denoted $\otimes$, the Hadamard (element-wise) product is denoted $*$ and the Khatri–Rao [9] product, which is the column-wise Kronecker product of two matrices, is denoted $\odot$. The use of these special products makes it possible to express most three-way models with two-way (matricized) arrays [10]. For example, a PARAFAC model of an $I \times J \times K$ array $\underline{\mathbf{X}}$ can be expressed as

$$\mathbf{X}^{(I \times JK)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} + \mathbf{E}^{(I \times JK)} \tag{12}$$

where $\mathbf{A}$ $(I \times R)$, $\mathbf{B}$ $(J \times R)$ and $\mathbf{C}$ $(K \times R)$ are component matrices and $\underline{\mathbf{E}}$ holds the residual unexplained variation. This notation is equivalent to the model

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}$$
$$i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K \tag{13}$$

2. TWO-WAY CENTERING

In order to understand when and how centering works, it is important first to consider the goals of centering and to realize how these goals are achieved in practice. These aspects are described in this section.

2.1. Reasons for centering

As stated by Harshman and Lundy [1], quite subjective and qualitative reasons are often given for performing centering. It is possible to formulate rational reasons for centering on scientific grounds. Basically, centering should be performed only if there are common offsets in the data or if modeling

such offsets provides an approximately reasonable model. Thus centering is performed to make interval-scale data behave as ratio-scale data, which is the type of data assumed in most multivariate models. Said more simply, centering should make a difference. This difference can manifest itself as:

(i) reduced rank of the model;
(ii) increased fit to the data;
(iii) specific removal of offsets;
(iv) avoidance of numerical problems.

Re (i). If a model of the raw data requires, say, $R+1$ components to describe the data well, whereas a model of the centered data requires only R components, then centering is sensible, because the model of the centered data only has $R(I+J)+J$ parameters. The J parameters pertain to the calculated averages, assuming that centering is performed across the first mode. The alternative of fitting the $(R+1)$-component model to the raw data would lead to a model with $(R+1)(I+J)$ parameters and thus would violate the parsimony principle. Re (ii). In some situations the rank of the appropriate model is not reduced upon centering, but if the fit of the model of the centered data is significantly improved, then naturally introducing extra parameters is useful. It is possible to heuristically consider the offsets introduced by centering as one extra 'half' component of which either the scores (centering across first mode) or the loadings (centering across second mode) are known *a priori* to be equal to one. This holds in the sense that the fit of a model with R components is poorer than the fit of a model with R components *and* offsets, which again is poorer than the bit of a model with $R+1$ components. Re (iii). Centering can remove certain offsets. In some situations the offsets as such are of interest, in which case it is interesting to estimate these. This can usually not be achieved by centering. Re (iv). In certain algorithms it may be useful to center the data in order to minimize algorithmic problems. For fitting a bilinear model using principal component analysis, it is known that the ratio of the two largest eigenvalues is related to the convergence rate of the power method (and related techniques such as NIPALS). For PARAFAC it is also known that if some components are strongly correlated, as evidenced through Tucker's congruence coefficient [11], then the fitting procedure may be complicated by so-called swamps. For both situations it holds that centering across certain modes can be helpful in minimizing the cause of the problem, because the resulting optimization problem is related to a different model with different (and hopefully better) properties with respect to numerical problems.

2.2. How centering works

2.2.1. *Centering can remove offsets because it is a projection step*

The following discussion pertains to centering across the first mode (ordinary column centering) but is readily applicable to centering across the second mode as well. Understanding that centering is a special projection step within one specific mode explains why it eliminates constant terms in the data (see Appendix I for details on projections).

If the vector $\mathbf{m}$ holds in its jth element the average of the jth column of $\mathbf{X}$, then $\mathbf{m}$ can be expressed as

$$\mathbf{m} = (1/I)\mathbf{X}^{\mathrm{T}}\mathbf{1} \Leftrightarrow \mathbf{m}^{\mathrm{T}} = (\mathbf{1}^{\mathrm{T}}/I)\mathbf{X} \tag{14}$$

where $\mathbf{1}$ is an I-vector of ones and $\mathbf{X}$ (the data) has size $I \times J$. Then centering $\mathbf{X}$ across the first mode (column centering) amounts to

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}\mathbf{m}^{\mathrm{T}} \tag{15}$$

where $\mathbf{Y}$ $(I \times J)$ contains the centered data. As

$$\mathbf{1}\mathbf{m}^{\mathrm{T}} = (\mathbf{1}\mathbf{1}^{\mathrm{T}}/I)\mathbf{X} \tag{16}$$

the centered data can also be expressed as

$$\begin{aligned}\mathbf{Y} &= \mathbf{X} - \mathbf{1}\mathbf{m}^{\mathrm{T}} = \mathbf{X} - (\mathbf{1}\mathbf{1}^{\mathrm{T}}/I)\mathbf{X} \\ &= [\mathbf{I} - (\mathbf{1}\mathbf{1}^{\mathrm{T}}/I)]\mathbf{X} = (\mathbf{I} - \mathbf{P})\mathbf{X} = \mathbf{P}^{\perp}\mathbf{X}\end{aligned} \tag{17}$$

where $\mathbf{PX} = \mathbf{P}[\mathbf{x}_1, \ldots, \mathbf{x}_J] = [\mathbf{P}\mathbf{x}_1, \ldots, \mathbf{P}\mathbf{x}_J]$, and $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. The matrix $(\mathbf{1}\mathbf{1}^{\mathrm{T}}/I) = \mathbf{P}$ is a symmetric and idempotent $(I \times I)$ matrix and is thus an (orthogonal) projection matrix (See Appendix I). This shows that the column averages are the orthogonal projections of the columns of $\mathbf{X}$ onto the direction of ones, i.e. the direction given by the vector $\mathbf{1}$. The centering matrix $\mathbf{I} - (\mathbf{1}\mathbf{1}^{\mathrm{T}}/I) = \mathbf{P}^{\perp}$ is the projection matrix onto the nullspace of $\mathbf{1}^{\mathrm{T}}$ which equals $range(\mathbf{1})^{\perp}$ (where $range(\cdot)$ is the range of a matrix). Stated otherwise, centering may also be interpreted as providing the residuals after regressing the columns of $\mathbf{X}$ onto $\mathbf{1}$. It follows that centering may be viewed as the projection of the data onto a space with the common offset (given by the I-vector $\mathbf{1}$) removed.

Mathematically, centering is a projection onto the nullspace of $\mathbf{1}^{\mathrm{T}}$ and it is worthwhile to keep this in mind. Suppose that the true model of the data contains offsets across the first mode. Then the model can be written as

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\mu^{\mathrm{T}} + \mathbf{E} \tag{18}$$

Projecting these data onto the nullspace of $\mathbf{1}^{\mathrm{T}}$ leads to

$$\mathbf{P}^{\perp}\mathbf{X} = \mathbf{P}^{\perp}\Phi\Theta^{\mathrm{T}} + \mathbf{P}^{\perp}\mathbf{1}\mu^{\mathrm{T}} + \mathbf{P}^{\perp}\mathbf{E} \Rightarrow \tag{19}$$

$$\mathbf{P}^{\perp}\mathbf{X} = \mathbf{Y} = \mathbf{P}^{\perp}\Phi\Theta^{\mathrm{T}} + \mathbf{P}^{\perp}\mathbf{E} \tag{20}$$

where $\mathbf{P}^{\perp}\mathbf{X} = \mathbf{Y}$ is the matrix holding the centered data, and the matrix $\mathbf{P}^{\perp}\mathbf{1}\mu^{\mathrm{T}}$ vanishes, as $\mathbf{1}$ has no residuals when projected onto itself. The part $\mathbf{P}^{\perp}\Phi\Theta^{\mathrm{T}}$ is a bilinear model with scores $\mathbf{P}^{\perp}\Phi$ and loadings Θ. Thus, instead of fitting the bilinear model *and* the offsets to the original data, it is only necessary to fit the bilinear model to the centered data $\mathbf{Y}$ with true structure $\mathbf{P}^{\perp}\Phi\Theta^{\mathrm{T}} + \mathbf{P}^{\perp}\mathbf{E}$. Centering also leads to models with residuals with zero column averages (centered across the first mode), because these are also projected onto the nullspace of $\mathbf{1}^{\mathrm{T}}$, as is clear from Equation (20).

If Φ and Θ are non-negative, then non-negativity constraints can be imposed on the model for $\mathbf{X}$. When $\mathbf{X}$ is centered, e.g. across the first mode, such a constraint is not meaningful for Φ anymore, because centering destroys the non-negativity of Φ (but not of Θ).

2.2.2. *Centering across several modes*

As mentioned earlier, centering across a given mode is called single centering. Single centering an array across one mode that has previously been single centered across another

mode is called double centering. Performing several single centerings (for multiway arrays, as many can be performed as the number of ways) is unproblematic, in the sense that centering across one mode leaves the 'centeredness' intact in other modes [1]. Further, the order of centering is immaterial. This means that if the data are first centered across the first mode, and subsequently the centered data are centered across the second mode, then the average of every column and every row will be zero. This follows because centering across the first mode can be written as $\mathbf{P}_I^{\perp}\mathbf{X}$, where $\mathbf{P}_I^{\perp}$ is the centering operator for the first mode. Centering across the second mode can be written as $\mathbf{X}\mathbf{P}_J^{\perp}$. Thus double centering can be written $\mathbf{P}_I^{\perp}\mathbf{X}\mathbf{P}_J^{\perp}$, and hence (i) the order in which centering is performed is immaterial and (ii) the double-centered array will have both column and row average zero, because $\mathbf{P}_I^{\perp}\mathbf{X}\mathbf{P}_J^{\perp}$ can be viewed as centering the matrix $\mathbf{X}\mathbf{P}_J^{\perp}$ across the first mode or centering the matrix $\mathbf{P}_I^{\perp}\mathbf{X}$ across the second mode.

2.2.3. *Centering is a two-stage procedure for a least squares fitting problem*

Consider a two-way data set which is generated as

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\mu^{\mathrm{T}} + \mathbf{E} \tag{21}$$

where Φ is $I \times R$, Θ is $J \times R$ and μ is $J \times 1$. It follows that the data can be modeled by a bilinear model plus a common offset for each variable/column plus additional unmodeled variation held in the residual matrix $\mathbf{E}$. This is the model assumed to be a valid approximation in most bilinear methods in chemometrics. The least squares loss function for the above model is

$$\|\mathbf{X} - (\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{n}^{\mathrm{T}})\|^2 \tag{22}$$

and this function is to be minimized directly over **A, B** and **n**, with **A, B** and **n** being of the same dimensions as Φ, Θ and μ respectively. The matrices **A, B** and **n** contain estimates of Φ, Θ and μ respectively, but it is intrinsic to the problem that these estimates do not uniquely recover the underlying parameters. For example, Φ and Θ can, at most, be found up to a rotation. As shown above, however, the parameters can be estimated in two steps. Centering the data across the first mode will remove the offsets μ, and the bilinear model is subsequently fitted to the centered data **Y**, thus minimizing the loss function

$$\|\mathbf{Y} - \mathbf{C}\mathbf{D}^{\mathrm{T}}\|^2 \tag{23}$$

only over **C** and **D** which are of the same dimensions as **A** and **B** above. It holds that

$$min\|\mathbf{X} - (\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{n}^{\mathrm{T}})\|^2 = min\|\mathbf{Y} - \mathbf{C}\mathbf{D}^{\mathrm{T}}\|^2 \tag{24}$$

i.e. the fit of the model fitted directly and the fit of the bilinear model fitted to the centered data will be exactly the same (see proof in Appendix II) even though the actual parameters will usually differ. This is an important result, because it guarantees the optimality of the model even if the offsets are calculated separately from the bilinear parameters.

The solution for minimizing Equation (22) directly is not unique for **n**. That is, the two-stage solution of centering first and then fitting the bilinear part gives one solution of many to the problem in Equation (22). Therefore centering removes μ, but **m** is not necessarily an estimate of μ.

This non-uniqueness is explained in short. Centering involves subtracting from each column its column average. The matrix

$$\mathbf{P}\mathbf{X} = \mathbf{P}\Phi\Theta^{\mathrm{T}} + \mathbf{P}\mathbf{1}\mu^{\mathrm{T}} + \mathbf{P}\mathbf{E} \tag{25}$$

holds in each row the vector $\mathbf{m}^{\mathrm{T}}$ containing the average value of each column of **X** (**P** is $\mathbf{1}\mathbf{1}^{\mathrm{T}}/I$, as above). If the part $\mathbf{P}\Phi\Theta^{\mathrm{T}}$ is a matrix of zeros, then **m** will be an estimate of the true offsets μ (with error **PE**), because

$$\mathbf{P}\mathbf{1}\mu^{\mathrm{T}} = \mathbf{1}\mu^{\mathrm{T}} \tag{26}$$

by definition. However, $\mathbf{P}\Phi\Theta^{\mathrm{T}}$ will only be a zero-matrix if Φ is orthogonal to **P** and/or $\Theta = \mathbf{0}$ (assuming that Φ and Θ have full column rank). That is, $\mathbf{P}\Phi = \mathbf{0}$ or $\Phi^{\mathrm{T}}\mathbf{P} = \mathbf{0}$. Thus the offsets will only equal the true offsets if the column space of Φ is orthogonal to **1** (meaning that Φ is centered already) or if $\Theta = \mathbf{0}$.

2.2.4. *Rank reduction and centering*

In some cases, centering reduces rank, and in some cases it does not. Column centering of **X** ($I \times J$) reduces the rank of **X** if and only if $\mathbf{1} \in range(\mathbf{X})$, where $range(\mathbf{X})$ is the range of **X** (see Reference [12], p. 156, and Reference [13]). Intuitively this is understandable. Centering is a projection. If the axis on which the projection takes place is a part of the range of **X**, then the residuals of this projection do not have this direction available anymore. Hence the rank of the matrix of residuals is lowered by one. This simple fact has several repercussions for centering across the first mode (column centering).

The following can be said about the noiseless case. Suppose that **X** ($I \times J$) is noiseless and can be modeled as

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\mu^{\mathrm{T}} = [\,\Phi \quad \mathbf{1}\,]\begin{bmatrix}\Theta^{\mathrm{T}} \\ \mu^{\mathrm{T}}\end{bmatrix} \tag{27}$$

where Φ is $I \times R$ of full rank and Θ is $J \times R$. Assume that

$$\begin{bmatrix}\Theta^{\mathrm{T}} \\ \mu^{\mathrm{T}}\end{bmatrix}$$

has full rank $R + 1$, which will be fulfilled for real data.

For **Y**, the column-centered **X**, two cases can be distinguished.

1. $\mathbf{1} \in range(\Phi) \Rightarrow rank(\mathbf{X}) = R \Rightarrow rank(\mathbf{Y}) = R - 1$.
2. $\mathbf{1} \notin range(\Phi) \Rightarrow rank(\mathbf{X}) = R + 1 \Rightarrow rank(\mathbf{Y}) = R$.

Hence in both cases the rank of **X** is reduced by one. The reverse also holds: if for the noiseless case no rank reduction of **X** is obtained upon column centering, then model (27) is not valid. To summarize, in the noiseless case there is a simple relation between the validity of model (27) and rank reduction upon column centering.

In the case with noise added, things are less simple. Suppose that **X** ($I \times J$) also contains noise and the model for **X** is

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\mu^{\mathrm{T}} + \mathbf{E} = [\,\Phi \quad \mathbf{1}\,]\begin{bmatrix}\Theta^{\mathrm{T}} \\ \mu^{\mathrm{T}}\end{bmatrix} + \mathbf{E} \tag{28}$$

If $I > J$, then in general $\mathbf{1} \notin range(\mathbf{X})$. Although $\mathbf{1} \in range$ ('noiseless' $\mathbf{X}$), this property is destroyed upon adding noise to $\mathbf{X}$. In the case $I \leq J$ and if $\mathbf{X}$ has full rank I, then $\mathbf{1}$ is by definition in the range of $\mathbf{X}$, whether or not there exists an offset term $\mathbf{1}\mu^{\mathrm{T}}$. Hence in the case with noise there is no simple relationship anymore between the validity of model (28) and mathematical rank reduction upon column centering.

2.3. When centering does not work

Viewing centering as a projection step rather than as a simple subtraction of averages has more than theoretical importance. In practice, situations often occur where subtraction of averages does not work and may in fact lead to models that fit the original data more poorly than if the data had not been preprocessed. This will be shown in the following for two different problems: handling missing data and modeling a single common offset.

2.3.1. Handling missing data

When data are missing, centering by subtracting averages from columns or rows does not lead to elimination of offsets. Rather, the offsets have to be eliminated simultaneously with the fitting of the bilinear part [14]. This is so because the equivalence between subtracting average values and projecting onto the nullspace of vectors of ones no longer holds, as the projection cannot be calculated with missing elements. As an example, consider the matrix $\mathbf{X}^{(1)}$ shown below:

$$\mathbf{X}^{(1)} = \begin{bmatrix} 10 & 20 \\ 7.5 & 15 \\ 1 & 2 \\ 1.5 & 3 \\ 0.5 & 1 \end{bmatrix}, \quad \mathbf{Y}^{(1)} = \begin{bmatrix} 5.9 & 11.8 \\ 3.4 & 7.8 \\ -3.1 & -6.2 \\ -2.6 & -5.2 \\ -3.6 & -7.2 \end{bmatrix}$$

$$\mathbf{X}^{(2)} = \begin{bmatrix} ? & 20 \\ ? & 15 \\ 1 & 2 \\ 1.5 & 3 \\ 0.5 & 1 \end{bmatrix}, \quad \mathbf{Y}^{(2)} = \begin{bmatrix} ? & 11.8 \\ ? & 7.8 \\ 0 & -6.2 \\ 0.5 & -5.2 \\ -0.5 & -7.2 \end{bmatrix} \tag{29}$$

This is a rank-one matrix and will remain so even after centering across the first mode. The averages of the two columns are 4.1 and 8.2 respectively and the centered matrix reads as $\mathbf{Y}^{(1)}$ in Equation (29), which is also a rank-one matrix.

Consider now a situation in which the first two elements in the first column are missing. The data then read as $\mathbf{X}^{(2)}$. This data set is naturally still perfectly modeled by a rank-one bilinear model, as no new information has been added. The averages of the two columns are now 1 and 8.2 respectively and subtracting these values leads to the centered matrix $\mathbf{Y}^{(2)}$. This matrix *cannot* be described by a rank-one model. This is easily realized by only looking at the last three rows. This is a rank-two submatrix, and the addition of the first two rows cannot change this. Thus, by subtracting averages from the data with missing elements, the structure of the data has been destroyed and meaningful results cannot be expected. Prior centering no longer leads to elimination of the true offsets as centering ordinarily does.

Centering is really an extension of the bilinear (or multi linear) model where offsets are assumed to be present in the model of the data. Data with missing elements constitute one situation in which such a model cannot be fitted in a least squares sense using centering. An alternative to eliminating offsets by preprocessing is given in Section 2.4.

2.3.2. Subtracting the grand mean

The traditional centering across the first mode easily leads to the belief that subtracting averages with the same structure as the offsets will generally eliminate these offsets. This holds for offsets constant across one mode, but it does not hold in general.

Consider a data set structured as a bilinear part plus a constant identical for all elements; that is, all elements have the same common offset, as also shown in Equation (4). It might seem natural to remove this offset by initially subtracting the grand mean m from the data. However, this will not simplify the subsequent modeling of the data, and, in fact, it obscures interpretation of the model, because such a centering leads to artificial mathematical components that also need to be modeled.

To explain this, assume that $\mathbf{X}$ is perfectly described by a bilinear part plus a common offset:

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu \tag{30}$$

Centering by removing the overall mean of all elements of $\mathbf{X}$ can be written as

$$\mathbf{Y} = \mathbf{X} - \mathbf{P}_I\mathbf{X}\mathbf{P}_J \tag{31}$$

where $\mathbf{P}_J$ is the projection matrix of $\mathbf{1}_J(=\mathbf{1}_J\mathbf{1}_J^{\mathrm{T}}/J)$ and $\mathbf{P}_I$ is the projection matrix of $\mathbf{1}_I(=\mathbf{1}_I\mathbf{1}_I^{\mathrm{T}}/I)$. Then $\mathbf{P}_I\mathbf{X}\mathbf{P}_J$ is a matrix of the same size as $\mathbf{X}$ holding the overall average of $\mathbf{X}$ in all its elements. Inserting the true model of Equation (30) in Equation (31) leads to

$$\begin{aligned} \mathbf{Y} &= \Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu - \mathbf{P}_I(\Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu)\mathbf{P}_J \\ &= \Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu - \mathbf{P}_I\Phi\Theta^{\mathrm{T}}\mathbf{P}_J - \mathbf{P}_I\mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu\mathbf{P}_J \\ &= \Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu - \mathbf{P}_I\Phi\Theta^{\mathrm{T}}\mathbf{P}_J - \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu \\ &= \Phi\Theta^{\mathrm{T}} - \mathbf{P}_I\Phi\Theta^{\mathrm{T}}\mathbf{P}_J \\ &= \Phi\Theta^{\mathrm{T}} - \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}s \end{aligned} \tag{32}$$

The scalar s is the overall average of the true bilinear part $\Phi\Theta^{\mathrm{T}}$. Even though the overall mean μ has been removed, a new common offset s has been introduced into the preprocessed data, and hence the same number of components is still necessary for modeling the data. Depending on the true parameters in the underlying model ($\Phi\Theta^{\mathrm{T}}$), the model fitted to the preprocessed data may therefore explain less or more of the original data than if the data had not been preprocessed! Clearly, preprocessing the data by subtracting the overall mean is generally *not* useful.

As subtracting the overall level does not remove the offset, another approach must apparently be adopted for handling situations with one common offset. There are basically two

Table I. Percentage of variation explained for different models of raw (left) and corrected (right) data

#LV	Raw data	Overall average subtracted
1	99.19%	98.05%
2	99.58%	99.08%
3	99.95%	99.67%
4	100.00%	100.00%
5	100.00%	100.00%

different ways of treating the problem. The best way is to optimize the loss function of the problem directly in a one-step procedure, rather than trying to use a two-step procedure where the offsets are first removed (see Section 2.4.).

Another simpler way of dealing with a constant offset follows from the observation that the model

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu \tag{33}$$

may equivalently be written as

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mu^{\mathrm{T}} \tag{34}$$

where

$$\mu = \mathbf{1}_J\mu \tag{35}$$

Posed in this way, it is evident that a model with one global offset is a special case of the situation treated earlier where each variable (or sample) has a specific offset. Therefore the constant offset may be eliminated by using ordinary centering across the first mode. As the offset is constant across rows, this centering removes the constant. An alternative procedure is to center across columns instead of rows, because the offset is also constant across columns.

An example is given for illustration. Consider a spectral data set. The data have been synthesized according to a bilinear three-component part plus a scalar offset of one. Thus the model is

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\mathbf{1}^{\mathrm{T}} \tag{36}$$

No noise is added. Using principal component analysis to model the raw data, four components are needed for describing the variation as expected (Table I, left). Modeling the data centered by subtracting the overall mean leads to a situation where four components still have to be used for describing all systematic variation (Table I, right). In fact, the three-component model explains less of the original data after preprocessing in this case.

Even though only three systematic components should be present, the loading plot (Figure 2, left) clearly shows that the first four components are 'spectral'-like. With proper preprocessing, only three loading vectors will be systematic, as shown in Figure 2 (right), using centering across the first mode.

Single centering involves fitting several parameters (I or J respectively). When there is only one constant parameter in the true model, as is the case here, a risk of overfitting is introduced with this approach. It is advisable, therefore, to center across the mode of largest dimension so that as few offsets as possible need to be estimated.

To recapitulate, the following rule establishes the 'correct' procedure for removing offsets before fitting the model: centering across one mode removes offsets constant across that mode *as well as* offsets constant across both modes [1]. This important rule also extends to multiway data of arbitrary order. Thus centering across one mode removes offsets constant across that mode *as well as* offsets constant across *several* modes involving that mode. This generalization follows from realizing that multiway models can

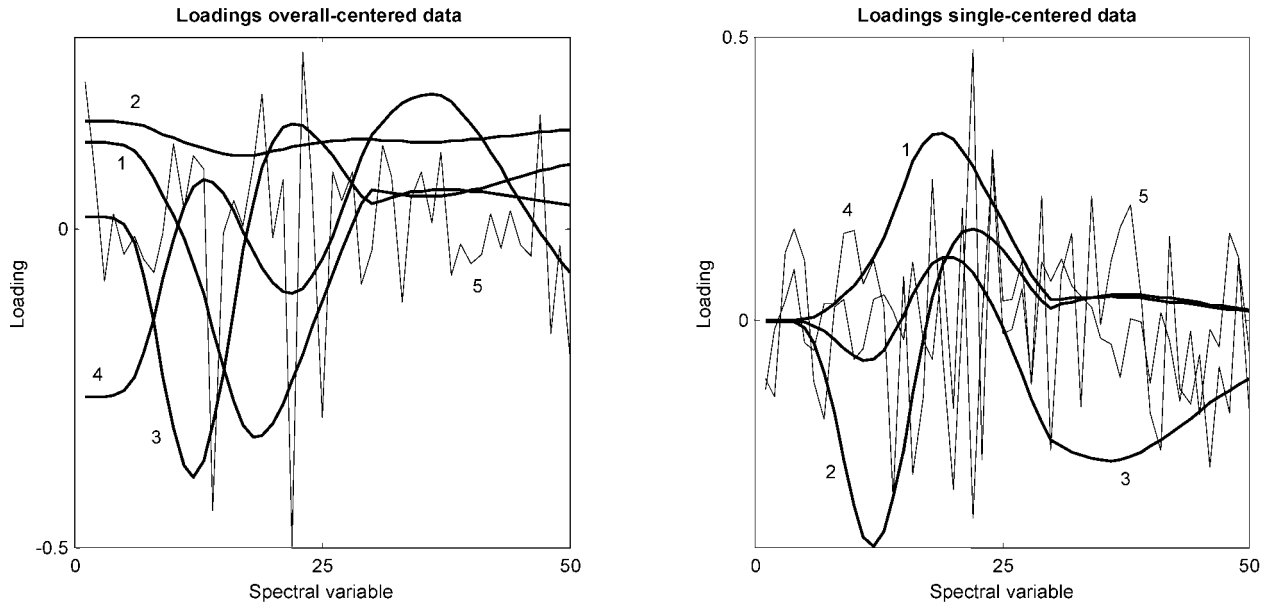

Figure 2. *Left.* Loading plot from principal component model of the data matrix where the overall average is subtracted. *Right.* The first five loadings are shown when the data have been correctly centered across the first mode.

always be considered as a constrained version of a bilinear model. Hence offsets constant across an arbitrary number of modes can always be considered a constrained version of a model with offsets constant across one mode. Centering across one of these modes will therefore eliminate the offsets because of the projection properties of centering.

2.4. Alternatives to centering

Instead of modeling the data in two steps—removing offsets by centering and then fitting a model to the residuals—it is possible to fit the model in one step, alleviating the need for projecting the offsets away. Two examples are given.

The example of missing data (Section 2.3.1) can be fitted directly in the following way. Assume that the offsets are, for instance, constant across the first mode and that a principal component analysis model is sought including offsets across the first mode. Such a PCA model of the data held in the matrix $\mathbf{X}$ including offsets reads as

$$\mathbf{X} = \Phi\Theta^{\mathrm{T}} + \mathbf{1}\mu^{\mathrm{T}} + \mathbf{E} \tag{37}$$

where μ is a J-vector. A very simple way to fit this model to a data matrix with missing elements in a least squares sense is by the use of an alternating least squares approach where the missing elements are continuously exchanged with their model estimates. Such an algorithm may proceed as follows.

1. Initialize missing values with reasonable values. Then the data set is complete and can be modeled by ordinary techniques.
2. Fit the model including offsets to the (now complete) data set. For PCA this amounts to centering the data and fitting the PCA model.
3. Exchange missing values in the data matrix for model estimates. These estimates will improve the current estimates and thus provide a data set where the estimated missing elements are closer to the correct ones according to the (yet unknown) true least squares model.
4. Proceed from step 2 until convergence.

This approach can be shown to converge, because it is an alternating least squares algorithm and hence has a non-increasing loss function. Upon convergence to the global minimum the imputed missing data will have no residuals and hence no influence on the model. The model parameters computed from the complete data are exactly those that would have been obtained had the model only been fitted to the non-missing data directly*. This approach can be viewed as a simple special case of expectation maximization [15]. For specific models or specific offsets, other approaches can also be feasible, but the above approach is general and easily implemented.

The problem with one common offset (Section 2.3.2) can be dealt with in the following way. The loss function for a bilinear model with constant overall offset is expressed as

$$\|\mathbf{X} - \Phi\Theta^{\mathrm{T}} - \mathbf{1}_I\mathbf{1}_J^{\mathrm{T}}\mu\|^2 \tag{38}$$

Instead of fitting the overparametrized model

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{E} \tag{39}$$

in a two-step procedure (see Equation (34)), it is possible to fit a 'correct' structural model

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} + m\mathbf{1}\mathbf{1}^{\mathrm{T}} + \mathbf{E} \tag{40}$$

directly. Instead of J parameters in $\mathbf{m}$, only one parameter m has to be estimated. The PCA model is used here as an example, but it may be exchanged with any other structural model, including a three-way model. Also, other types of offsets may be used. The loss function may be optimized in several ways leading to a least squares model [16,17]. A simple algorithm is based on alternating least squares where first the offset is set to an initial value. Then the structural model is fitted to the corrected data $\mathbf{X} - m\mathbf{1}\mathbf{1}^{\mathrm{T}}$ using an ordinary PCA algorithm in this case. This provides an update of the bilinear parameters. Subtracting the new interim PCA model from the data leads to

$$\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} = m\mathbf{1}\mathbf{1}^{\mathrm{T}} + \mathbf{E} \tag{41}$$

Therefore m may be determined conditional on $\mathbf{A}$ and $\mathbf{B}$ as the overall average of $\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}}$. By alternating between updating the loading parameters and the offset, the model parameters will be estimated upon convergence.

In a three-way context using a PARAFAC model, an auxiliary benefit is that the offsets may often be uniquely determined owing to the uniqueness of the PARAFAC model.

2.5. Summary

Offsets are part of the model hypothesized for the data. Some offsets can be removed by centering before fitting the remaining part of the model. This removal of offsets involves fitting additional parameters. *Proper centering* is defined as a centering operation that removes the offsets postulated and does not change the structural model of the data. Proper centering is always performed across a single mode. Sequential centering across several modes is allowed. Hence proper centering can always be written as $\mathbf{P}_I^{\perp}\mathbf{X}, \mathbf{X}\mathbf{P}_J^{\perp}$ or $\mathbf{P}_I^{\perp}\mathbf{X}\mathbf{P}_J^{\perp}$, depending on whether centering is performed across the first, second or both modes. Here $\mathbf{P}_I^{\perp}$ is an $I \times I$ matrix defined as $\mathbf{I} - (\mathbf{1}\mathbf{1}^{\mathrm{T}}/I)$, where $\mathbf{I}$ is an $I \times I$ identity matrix and $\mathbf{1}$ is an I-vector of ones. The matrix $\mathbf{P}_J^{\perp}$ is defined similarly. If elements are missing or other offsets are to be modeled, this has to be done using a one-step modeling approach where offsets and other parameters are considered simultaneously (usually using iterative algorithms).

3. TWO-WAY SCALING

Unlike centering, scaling does not change the structure of the model. Scaling is used to change the weights given to different parts of the data in fitting the model. Although scaling is important, it usually has a much less dramatic influence on the fitted model than centering, as long as the

* If the algorithm for fitting the bilinear part of the model is iterative, it is useful not to iterate until convergence in each step 2. Instead, only a few iterations are performed before one proceeds to step 3 where the complete model ($\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{m}^{\mathrm{T}}$ in the case of PCA) is calculated and used for obtaining better estimates of the missing values.

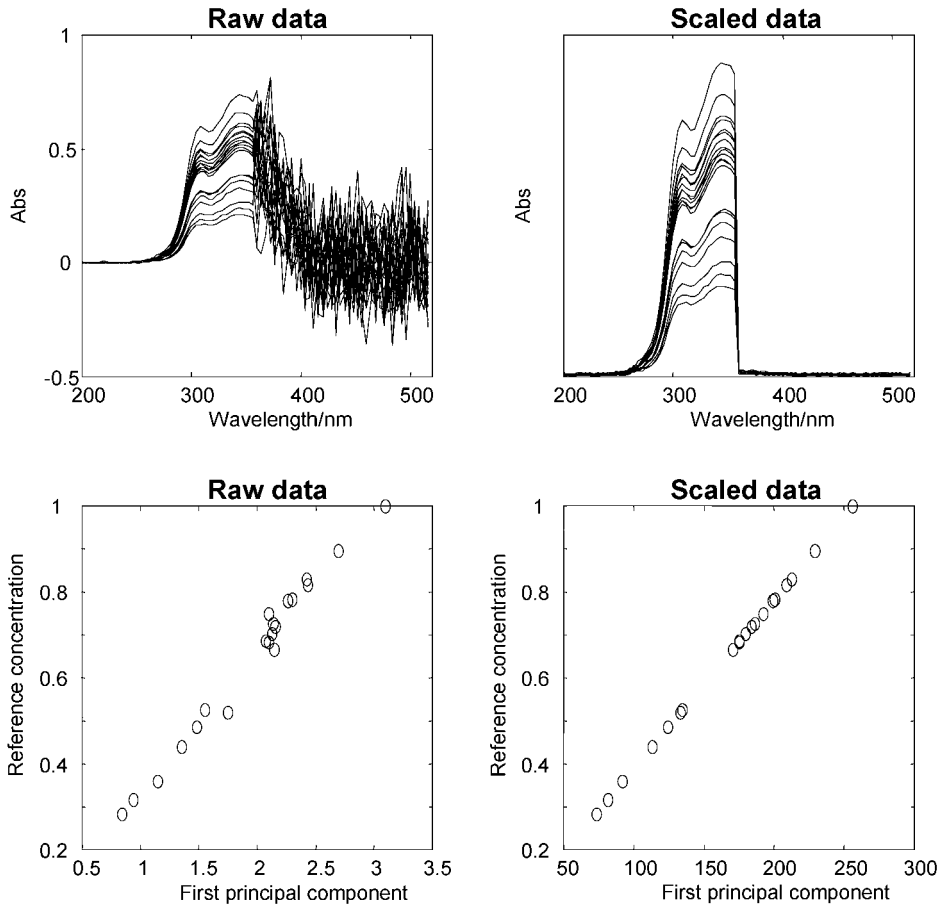

Figure 3. Influence of scaling on fitted model.

model and scaling are reasonable [18]. Some issues related to scaling are discussed by Paatero and Tapper [19].

3.1. Reasons for scaling

Scaling is used for several reasons. Some important ones are:

(i) to adjust scale differences;
(ii) to accommodate for heteroscedasticity;
(iii) to allow for different sizes of subsets of data (block scaling).

Re (i). It is quite common to use, for instance, autoscaling (centering across the first mode and scaling to unit standard deviation within the second mode) to let the variance of each variable be identical initially. Thereby all variables have the same variance, and as the subsequent fitting of a model is performed so as to describe as much systematic variation as possible, every variable has the same initial opportunity of entering the model. This type of scaling is especially useful when the variables are measured in different measurement units (e.g. Pa, °C, …) Re (ii). The ordinary least squares fitting of a model is statistically optimal in a maximum likelihood sense if the errors are homoscedastic, independent and Gaussian. If the variances of the distributions are not the same, though the same within e.g. a specific variable, it is possible to accommodate the fitting procedure accordingly by initially scaling the data within the variable mode. By scaling each variable with the inverse of the standard deviation of the residual variance, the fitted model will be optimal in a maximum likelihood sense. Re (iii). When the data are made up of several subsets of very different sizes, it may sometimes be advantageous to scale each block separately in order to ensure that all the different blocks are allowed to influence the model. Consider for example a situation in which 5000 wavelengths are measured in an infrared spectrum (absorbance between 0 and 1), and one variable is given for the temperature. Owing to the huge difference in number of variables (5000 and one respectively), the total variance of the infrared spectra will be tremendous compared with that of the temperature. If no scaling is applied to adjust for this difference, then the model is implicitly forced to focus on the infrared data. Explaining the temperature variable will not lead to a well-fitting model, unless the model is so complex that it can fit both subsets simultaneously or the temperature data are in accordance with the infrared data. If the infrared data and the temperature data are initially believed to be equally important, then scaling both subsets to the same *total* variance will provide a model that reflects this assumption. Thus, in this case, scaling is used from an information point of view to ensure that all important information can enter the model, irrespective of the variance of the different sources of information.

It is important to note that even if no scaling is applied, the data are still scaled by the weight one. Thus scaling (or the

loss function in general, as will be shown in the next section) always has to be considered before fitting the model.

All the above-mentioned reasons can be put under the same heading by the term weighted least squares fitting, which is a general and broader approach to fitting models than merely the use of scaling. This will be elaborated on in the next section.

3.2. How scaling works

Scaling is a subject often treated in conjunction with centering. However, the purpose of scaling is very different from that of centering. Scaling is a way of introducing a loss function other than the least squares loss function normally used. Therefore scaling *does not* change the interpretation of the model and its parameters. As for centering, scaling has to be performed in a specific way in order not to introduce artificial structure that needs to be modeled. This becomes even more apparent when going to three-way models.

3.2.1. Different types of scaling

Scaling is usually performed by multiplying each column or each row in the data matrix by a scalar. There are two types of scaling that are relevant for two-way matrices. One is scaling within the first mode, where every row is multiplied by a specific number:

$$\mathbf{Y} = \mathbf{W}\mathbf{X} \tag{42}$$

where $\mathbf{W}$ is an $I \times I$ diagonal matrix with the scaling parameter for the ith row on its ith diagonal element. This is the type of scaling used for example in standard normal variate correction [20–22], where the norm or area of each row of $\mathbf{X}$ is scaled to the same scalar value by using an appropriate $\mathbf{W}$. It extends easily to multiway arrays, as discussed in Section 5.2.

The other main type of scaling is within the second mode, where every column is multiplied by a specific number:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} \tag{43}$$

where $\mathbf{W}$ is here a $J \times J$ diagonal matrix with the scaling parameter for the jth column on its jth diagonal element. This is the scaling ordinarily used in e.g. PCA, where the weight of a specific column (variable) is often chosen to be the inverse of the standard deviation of the variable. In combination with centering across the first mode, such scaling within the second mode is often referred to as autoscaling in the two-way case.

An example of scaling according to Equation (43) is shown in Figure 3. One-component bilinear data with huge random residual variation in the last half of the variables (upper left) are generated. The resulting $\mathbf{X}$ has the spectra (as plotted in Figure 3, upper left) in its rows. The first principal component is seen to correlate well with the reference score generating the data (lower left). However, when the data are initially weighted as in Equation (43) by the inverse of the residual standard deviation (upper right), the right part of the data is downweighted substantially. The correlation between the true known score ('concentration') and the estimated score of the scaled data is even higher (lower right) than for the raw data (left). Even in this extreme case the influence of the noise is not at all as drastic as might be expected. This illustrates that scaling is not as critical as centering, as long as the scaling is reasonable and the variables are relevant. Note that the scaling in this example is not the usual autoscaling using the inverse standard deviation of the data. Rather, the inverse standard deviation of the residual variation, for instance, assessed by replicates, is used.

3.2.2. Scaling and weighted least squares fitting

Given a data matrix $\mathbf{X}$ ($I \times J$) and a model $\hat{\mathbf{X}}$ ($I \times J$), which may for example be a bilinear model ($\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^{\mathrm{T}}$), a standard approach for determining the model and its parameters is to fit the model in a least squares sense by minimizing the loss function

$$\|\mathbf{X} - \hat{\mathbf{X}}\|^2 \tag{44}$$

which can also be expressed as

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \hat{x}_{ij})^2 \tag{45}$$

When the different elements of the data have different uncertainties or relevances, it is possible to fit the model using a weighted least squares loss function. In one of its simplest forms this can be expressed as

$$\|(\mathbf{X} - \hat{\mathbf{X}}) * \widetilde{\mathbf{W}}\|^2 \tag{46}$$

where $\widetilde{\mathbf{W}}$ ($I \times J$) holds in its ijth element the weight of the ijth element of $\mathbf{X}$, and * denotes the Hadamard (element-wise) product. Often the weight of an element is set equal to the inverse of the standard deviation of the residual variation. It is also possible to use more elaborate weights if there are certain correlations between the residuals [23,24]. The above weighted loss function can also be expressed in scalar notation as

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \hat{x}_{ij})^2\tilde{w}_{ij}^2 \tag{47}$$

In a maximum likelihood sense this loss function is optimal if the weights reflect the uncertainty of the individual elements, if there is no correlation between the residuals of different elements and if the residuals are normally distributed. If the uncertainty of a given variable is the same over all objects, the above model turns into

$$\|(\mathbf{X} - \hat{\mathbf{X}}) * \widetilde{\mathbf{W}}\|^2 = \|(\mathbf{X} - \hat{\mathbf{X}})\mathbf{W}\|^2 \tag{48}$$

where the weight matrix $\mathbf{W}$ is now a diagonal matrix which holds the column-specific weights in the diagonal. The loss function may be transformed as follows:

$$\|(\mathbf{X} - \hat{\mathbf{X}})\mathbf{W}\|^2 = \|\mathbf{X}\mathbf{W} - \hat{\mathbf{X}}\mathbf{W}\|^2 \tag{49}$$

In the case of a bilinear model the above reads as

$$\|(\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}})\mathbf{W}\|^2 = \|\mathbf{X}\mathbf{W} - \mathbf{A}\mathbf{B}^{\mathrm{T}}\mathbf{W}\|^2 = \|\mathbf{X}\mathbf{W} - \mathbf{A}\mathbf{H}^{\mathrm{T}}\|^2 \tag{50}$$

Thus, by fitting the bilinear model $\mathbf{A}\mathbf{H}^{\mathrm{T}}$ to the data scaled within the second mode, $\mathbf{X}\mathbf{W}$, *in an ordinary least squares sense*, the weighted loss function of Equation (48) is automatically optimized. This is the basic mathematical rationale behind scaling. If the sought model $\hat{\mathbf{X}}$ has the structure $\mathbf{A}\mathbf{B}^{\mathrm{T}}$, then

fitting a bilinear model to the scaled data provides the model in the form $\mathbf{AH}^{\mathrm{T}}$. Thus the score matrix (or a rotated version of it) is directly provided in $\mathbf{A}$, whereas the loadings of the problem are found by premultiplying the found loadings $\mathbf{H}$ by $\mathbf{W}^{-1}$, as

$$\mathbf{B}^{\mathrm{T}}\mathbf{W} = \mathbf{H}^{\mathrm{T}} \Rightarrow \mathbf{B}^{\mathrm{T}} = \mathbf{H}^{\mathrm{T}}\mathbf{W}^{-1} \Rightarrow \mathbf{B} = \mathbf{W}^{-1}\mathbf{H} \qquad (51)$$

Sometimes, only the scaled data are considered and the model parameters are not transformed back to the original domain. The appropriateness of this approach is still, however, governed by the fact that scaling as outlined above maintains the bilinear structure assumed reasonable for the raw data.

There is a direct connection between $\|(\mathbf{X} - \mathbf{AB}^{\mathrm{T}})\mathbf{W}\|^2$ and $\|\mathbf{XW} - \mathbf{AH}^{\mathrm{T}}\|^2$. However, there is no direct connection between the solution to $\|(\mathbf{X} - \mathbf{AB}^{\mathrm{T}})\mathbf{W}\|^2$ and $\|(\mathbf{X} - \mathbf{TP}^{\mathrm{T}})\|^2$ unless the model has perfect fit. That is, fitting the bilinear model to scaled and unscaled data represents two different problems with no direct relation.

3.3. When scaling does not work

When scaling within several modes is desired, the situation is complicated, because scaling one mode affects the scale of the other mode. For example, scaling to a standard deviation of one within both the first and second modes will generally not be possible, not even using iterative scaling [1]. If scaling to a mean square of one is desired within both modes, this has to be done iteratively until convergence [1,4]. Using mean squares rather than, for instance, standard deviations for scaling has the attractive property that iterative scaling is guaranteed to converge in the case where no centering is included in the iterative scheme [1].

Iterative preprocessing may seem unsatisfactory, because it tends to complicate the subsequent evaluation and validation of the preprocessing, since more than one set of scaling parameters for each mode has to be used. These several matrices holding the scaling parameters from each iteration may be combined, though, into a single matrix [1]. An *ad hoc* alternative is to skip the iterative preprocessing and perform only one scaling of each mode. The purpose of scaling is mainly to bring the levels of variation of different variables to some sort of equivalent level. Therefore one iteration of scaling can suffice to scale the data, so that no part of the data will have an unreasonably large influence on the subsequent fitting.

3.4. Alternatives to scaling

As shown in the preceding paragraphs, scaling can be considered to be a special case of using a weighted least squares loss function. When more complicated weights are needed, it is not always possible to fit the model indirectly by fitting the least squares model to the scaled data. In such situations an alternative to scaling is to use algorithms that directly handle a weighted least squares optimization criterion [10,25,26]. This can be relevant, for example, when the residual variation is correlated across both rows and columns [23,27,28].

3.5. Summary

Scaling does not affect the structural model of the data, but the loss function used to estimate the model parameters. *Proper scaling* is defined as the scaling that can be expressed in terms of pre- or postmultiplied weights in the loss function of the model. Hence proper scaling of $\mathbf{X}$ can always be expressed as $\mathbf{W}_I\mathbf{X}$, $\mathbf{XW}_J$ or $\mathbf{W}_I\mathbf{XW}_J$, where $\mathbf{W}_I$ and $\mathbf{W}_J$ are diagonal matrices holding, for example, the inverse standard deviation of the corresponding row or column. If a certain scale is needed for both modes, then the corresponding weights have to be found iteratively. Scaling, for example, to unit standard deviation in two modes is not possible in general, whereas scaling to unit mean square variation is possible.

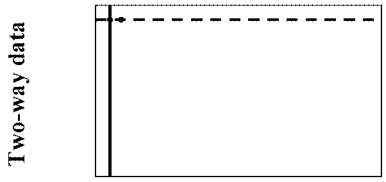

Figure 4. Two-way array showing the dependence between centering and scaling.

4. SIMULTANEOUS TWO-WAY CENTERING AND SCALING

A complicating issue in preprocessing is the interdependence of centering and scaling [29]. Because preprocessing is mostly performed in one or a few standardized ways in two-way analysis, the problems are seldom appreciated. It is important, however, to be aware that not all combinations of centering and scaling will work as anticipated (see e.g. Reference [1]). Generally, only centering across both modes is straightforward, or scaling within one mode combined with centering across the other mode [1,30], which is exactly what e.g. autoscaling amounts to.

4.1. Scaling within a mode disturbs centering across the same mode but not across other modes

Scaling within one mode disturbs prior centering across the same mode but not across other modes [4]. This holds for two-way arrays as well as higher-order arrays. The reason for this is illustrated in Figure 4. The full line shows a typical column vector and the broken line a typical row vector. When scaling within the first mode, the elements of any column are multiplied by different numbers, and hence prior centering across the first mode is destroyed.

Consider a two-way array $\mathbf{X}$ ($I \times J$). If the array is scaled within the first mode, this can be expressed as

$$\mathbf{Y} = \mathbf{WX} \qquad (52)$$

where $\mathbf{Y}$ is the scaled array, $\mathbf{W}$ is an $I \times I$ diagonal matrix holding the scaling parameters, and $\mathbf{X}$ is the original array. As can be seen, scaling within the first mode amounts to multiplication of every row by a scalar. This does not affect any centering of the vectors across the second mode, because every element in a row vector is multiplied by the same number. The average of any row will be the original average of the row scaled down accordingly, and therefore, if the

average is zero, it will stay zero. In the first mode, however, each element in a column is multiplied by a different scalar. If centering is performed across the first mode, these column vectors will not necessarily preserve their zero average after subsequent scaling within the first mode. Mathematically, the centered matrix $\mathbf{P}^{\perp}\mathbf{X}$ becomes $\mathbf{W}\mathbf{P}^{\perp}\mathbf{X}$ upon scaling. As

$$\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{P}^{\perp}\mathbf{X} \neq \mathbf{0},$$

the preprocessed matrix is no longer guaranteed to be centered. Offsets constant across the first mode, however, *will* still be removed, because

$$\mathbf{P}^{\perp}(\Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mu^{\mathrm{T}}) = \mathbf{P}^{\perp}\Phi\Theta^{\mathrm{T}} \Rightarrow \mathbf{W}\mathbf{P}^{\perp}(\Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mu^{\mathrm{T}}) = \mathbf{W}\mathbf{P}^{\perp}\Phi\Theta^{\mathrm{T}} \quad (53)$$

Note also the interesting fact that if scaling is performed before centering, the result will be different. In that case the original offsets will not be removed, but the data will be centered (yielding centered scores and residuals), because

$$\mathbf{P}^{\perp}\mathbf{W}(\Phi\Theta^{\mathrm{T}} + \mathbf{1}_I\mu^{\mathrm{T}}) = \mathbf{P}^{\perp}\mathbf{W}\Phi\Theta^{\mathrm{T}} + \mathbf{P}^{\perp}\mathbf{W}\mathbf{1}_I\mu^{\mathrm{T}} \neq \mathbf{P}^{\perp}\mathbf{W}\Phi\Theta^{\mathrm{T}} \quad (54)$$

This holds because, unlike for $\mathbf{P}^{\perp}\mathbf{1}$, it does not hold that $\mathbf{P}^{\perp}\mathbf{W}\mathbf{1}$ is a zero-vector in general.

4.2. Centering across one mode disturbs scaling within all modes

Centering across one mode disturbs scaling within all modes. This holds for two- as well as multiway arrays, but there are certain cases for which it does not hold. One of these special cases is the situation in which two-way data are scaled within the second mode to a standard deviation of one and subsequently centered across the first mode (autoscaling)*. This subsequent centering will not disturb the scaling within the second mode (though it would disturb scaling within the first mode had that been performed). The reason is that the scaling is specifically performed relative to the center of the data (standard deviations are based on centered data). Hence any change in offset is immaterial for the standard deviation. Scaling by means other than standard deviations will not have this property. In multiway analysis it is common to use mean squares for scaling instead of standard deviations, because such scaling more often converges when implemented in an iterative scheme, and because scaling by standard deviations implicitly assumes an offset, which may or may not be present depending on the structural part of the model.

4.3. Centering across and scaling within the same mode is problematic

Centering across a mode within which scaling is also applied, or *vice versa*, is generally not going to retain all the properties of the two individual operations, as discussed earlier. For example, if centering across the first mode and scaling within the first mode are desired, then setting

$$\mathbf{Y} = \mathbf{W}\mathbf{P}^{\perp}\mathbf{X} \quad (55)$$

* Normally, centering would be performed before scaling for computational reasons, as the averages are needed for scaling by the inverse standard deviation.

All elements have the same offset

$$\mathbf{X}_k = \Phi\mathbf{D}_k\Theta^{\mathrm{T}} + \lambda\,\mathbf{1}_I\mathbf{1}_J^{\mathrm{T}} + \mathbf{E}_k,\ k=1,..,K$$

$$\mathbf{X}^{(I\times JK)} = \Phi(\Omega\odot\Theta)^{\mathrm{T}} + \lambda\,\mathbf{1}_I(\mathbf{1}_K\odot\mathbf{1}_J)^{\mathrm{T}} + \mathbf{E}^{(I\times JK)}$$

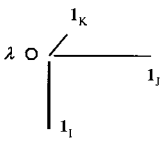

Figure 5. Structure of offsets in three-way (trilinear) data when all elements have the same offset held in the scalar λ. Two alternative but equivalent ways of writing the PARAFAC model are shown: the slab notation using submatrices $\mathbf{X}_k$, and the notation described in Equation (12).

with $\mathbf{W}$ being a diagonal scaling matrix and $\mathbf{P}^{\perp}$ a centering operator, will not lead to a preprocessed matrix in which the first mode vectors are centered, although possible offsets will be eliminated. Conversely, setting

$$\mathbf{Y} = \mathbf{P}^{\perp}\mathbf{W}\mathbf{X} \quad (56)$$

will not eliminate the original offsets, even though the preprocessed array will have centered first-mode vectors. Hence for a specific application a choice has to be made between these two approaches, depending on why the centering is applied.

4.4. Summary

Proper centering is a centering operation that correctly removes the presupposed offsets and does not introduce other offsets into the data. Likewise, *proper* scaling introduces the correct weights into the loss function. Stated otherwise, proper centering and scaling do what they are supposed to do. Proper centering and scaling can sometimes be combined. Unproblematic combinations can always be expressed as

$$(\mathbf{P}^{\perp}\mathbf{X})\mathbf{W} \text{ or } \mathbf{W}(\mathbf{X}\mathbf{P}^{\perp}) \quad (57)$$

where $\mathbf{P}^{\perp}$ is the projection matrix that centers the data, and $\mathbf{W}$ is a weighting matrix, both of appropriate size. The parentheses in Equation (57) indicate the proper order in which the different preprocessing steps have to be performed. If performed oppositely, offsets will not be removed, although the data will be centered. Problematic combinations are

$$\mathbf{W}\mathbf{P}^{\perp}\mathbf{X} \text{ or } \mathbf{X}\mathbf{P}^{\perp}\mathbf{W} \quad (58)$$

These combinations do not retain all the properties desired of the preprocessing steps.

5. THREE-WAY PREPROCESSING

The preprocessing of multiway arrays will now be discussed using three-way arrays as an example. The basic properties discussed thus far are unchanged. Centering has to be performed across a specific mode, and scaling has to be performed by a transformation within a specific mode. Most difficulties in preprocessing three-way arrays arise because of the problems outlined so far, which all generalize to multiway arrays. The problems are sometimes enhanced, because three-way data are often rearranged (matricized) to

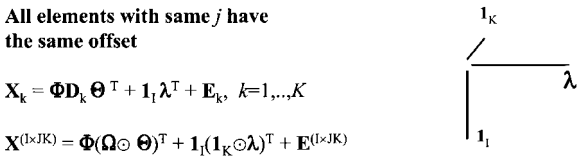

Figure 6. Structure of offsets in three-way (trilinear) data when all elements in each vertical slab have the same offset. The offsets are held in the *J*-vector λ.

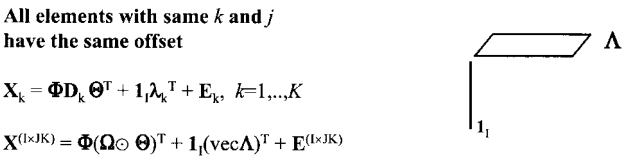

Figure 7. Structure of offsets in three-way (trilinear) data when all elements in each vector have the same offset (case three). The offsets are held in a matrix Λ ($J \times K$) whose *jk*th element holds the offset of the vertical column with second- and third-mode indices *j* and *k*. The vector λ_k is the *k*th row of the matrix Λ.

two-way arrays before preprocessing. This is unfortunate, because it introduces a column mode that is a combination of two of the original modes. Transformation within or across this combined mode should be avoided if multiway models are to be fitted, because the mode is not a 'real' mode but merely a computational construct.

5.1. Centering

5.1.1. Possible three-way offsets and their proper removal

The observations on centering of two-way data are helpful in discussing centering of three- and higher-way arrays. If the basis of two-way centering is understood, then three-way centering is quite simple. In the following it will be assumed that the true model of the data is a PARAFAC model plus possible offsets, but the conclusions hold for any multilinear model.

Consider a three-way array. Conceptually, offsets may occur in three different ways, i.e. constant across all modes (Figure 5), constant across two modes (Figure 6) or constant across one mode only (Figure 7). In the figures the first-mode loadings are held in the $I \times R$ matrix Φ, the second-mode loadings in the $J \times R$ matrix Θ and the third-mode loadings in the $K \times R$ matrix Ω. The matrix $\mathbf{D}_k$ is an $R \times R$ diagonal matrix holding the *k*th row of Ω in its diagonal.

Regardless of the structure of the offsets, the basic principle of centering is that the data must be preprocessed, so that they are projected onto the nullspace of vectors of ones in a particular mode. For the first mode, projecting a data array $\underline{\mathbf{X}}$ onto the nullspace of $\mathbf{1}^T$, i.e. centering across the first mode, amounts to

$$\mathbf{Y} = \mathbf{P}^{\perp}\mathbf{X}^{(I\times JK)} \tag{59}$$

where $\mathbf{P}^{\perp} = \mathbf{I} - (\mathbf{1}\mathbf{1}^T/I)$. For the second and third modes the centering can be performed similarly. As mentioned earlier, such centering is referred to as *single centering*. Centering, for example, across the first mode of an array can thus be done by matricizing the array to an $I \times JK$ matrix and then centering this matrix across the first mode as in ordinary two-way analysis:

$$y_{ijk} = x_{ijk} - \frac{\sum_{i=1}^{I} x_{ijk}}{I} \tag{60}$$

The column mean is subtracted from every element, as depicted graphically in Figure 8. As can be seen, single centering is similar in structure to the type of offsets shown in Figure 7.

Such single centerings performed successively across two modes are referred to as double centering. That is, double centering is performed by first centering across one mode and then centering the outcome across another mode. The order of centerings is immaterial, but it is essential that they are performed sequentially. For all three situations depicted in Figures 5–7 the above centering across the first mode will remove the shown offsets, because in all three situations the offsets are constant across the first mode.

As for the two-way case, *only* single centering leads to the properties sought in centering (removal of offsets). Other types of centering, such as subtracting the overall mean, will introduce artifacts that have to be modeled additionally to the inherent systematic variation. This was shown for the two-way case in Section 2.3.2. Such types of incorrect centering are often used in three-way analysis. For example, matricized data are centered across a combined mode. For

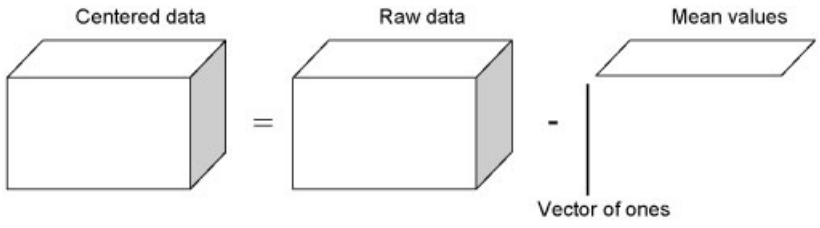

Figure 8. Centering across the first mode. For each column of the raw data a mean value is calculated and subtracted from each element in the column. Thus a two-way matrix of mean values is obtained. Note that this centering is identical to matricizing the data to a two-way structure and centering these two-way data across the first mode.

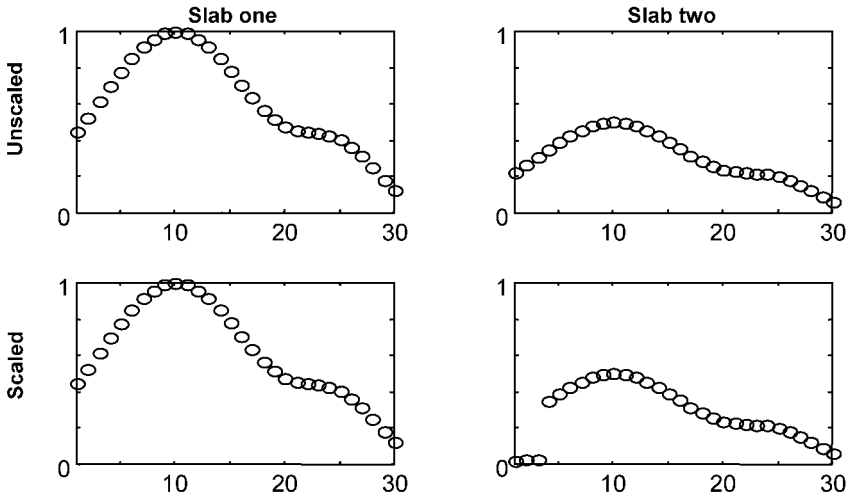

Figure 9. Incorrect scaling of three-way array. Spectra from one sample measured at two conditions (slabs one and two). The top plot shows the data before scaling and the lower plot after a hypothetical scaling.

example, if an array of structure (variables × time × samples) is centered by subtracting the average calculated across samples *and* time, then, in line with the above example, artificial offsets are introduced and the subsequent model will have to fit this additional variation as well. This can obscure validation and exploration of the model and will lead to models that do not provide overall least squares solutions.

5.2. Scaling

As explained for the two-way case, scaling is a transformation of a particular variable (or object) space. Instead of fitting the model to the original data, the model is fitted to the data transformed by a (usually) diagonal scaling matrix in the mode whose variables are to be scaled. This means that whole matrices instead of columns have to be scaled by the same value in three-way analysis. For a four-way array, three-way slabs would have to be scaled by the same scalar. Mathematically, scaling *within the first mode* can be described as

$$y_{ijk} = w_i x_{ijk} \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K \quad (61)$$

where $\underline{\mathbf{Y}}$ with elements y_{ijk} is the scaled array and, for instance, setting

$$w_i = \frac{1}{\sqrt{\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{x_{ijk}^2}{JK}}} \quad (62)$$

will scale to a unit mean square within the sample mode. Using matricized arrays, scaling may be expressed as

$$\mathbf{Y}^{(I\times JK)} = \mathbf{W}\mathbf{X}^{(I\times JK)} \quad (63)$$

where $\mathbf{W}$ is an $I \times I$ diagonal matrix holding the scaling values in its diagonal. The assumed structural model of the data is a multilinear model, e.g. a PARAFAC model $\hat{\mathbf{X}} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}}$. With the above scaling, a similar structural model, $\hat{\mathbf{Y}} = \mathbf{P}(\mathbf{R} \odot \mathbf{Q})^{\mathrm{T}}$, will hold for the transformed data, because

$$\begin{aligned}\|\mathbf{Y} - \mathbf{P}(\mathbf{R} \odot \mathbf{Q})^{\mathrm{T}}\|^2 &= \|\mathbf{W}\mathbf{X} - \mathbf{P}(\mathbf{R} \odot \mathbf{Q})^{\mathrm{T}}\|^2 \\ &= \|\mathbf{W}(\mathbf{X} - \mathbf{W}^{-1}\mathbf{P}(\mathbf{R} \odot \mathbf{Q})^{\mathrm{T}})\|^2 \quad (64)\end{aligned}$$

The loss function minimized when fitting a model with a weighted criterion is

$$\|\mathbf{W}(\mathbf{X} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}})\|^2 \quad (65)$$

and hence it holds that the parameters of this model can be found by fitting the scaled data and setting

$$\mathbf{A} = \mathbf{W}^{-1}\mathbf{P}, \qquad \mathbf{B} = \mathbf{Q}, \qquad \mathbf{C} = \mathbf{R} \quad (66)$$

Thus fitting the scaled data provides a solution not only to the problem posed as fitting a model to $\mathbf{Y}$, but to the problem of fitting a model to $\mathbf{X}$ and where the first-mode loadings obtained are transformed by the scaling matrix $\mathbf{W}$. As for two-way scaling, it is emphasized that the found model parameters have no direct relations to the parameters found when fitting the model to the raw data in a least squares sense.

5.2.1. *Incorrect scaling of three-way arrays*

Scaling has to be applied by transforming the data within a given mode. It is not appropriate to scale an array within two combined modes, which can happen, for example, when autoscaling a matricized array. Such an inappropriate scaling will lead to the inclusion of artificial components in the data.

Consider an $I \times 30 \times 2$ three-way array with I samples, 30 variables (say spectral), measured at two conditions (e.g. two different pH values) as shown in Figure 9. In the two top plots the profiles of the first hypothetical sample are shown (30 variables). To the left the measured spectrum is shown at the first condition and to the right at the second condition. As can be seen, the shape is identical in the two plots; only a multiplicative factor distinguishes the profiles. In this case, only one component is necessary for describing this

variation. If the array is scaled such that different occasions of the same variable are scaled differently, then the phenomena can no longer be described by the variation of one basic profile. This is shown in the lower plots, where the first three wavelength variables have been scaled differently in slabs one and two. In slab one they have been scaled by one and in slab two by 0.01. It is clear that to preserve the multilinearity of the data, any occurrence of a given variable must be scaled by the same amount.

To show the influence of incorrect scaling, consider a synthetic data set with PARAFAC structure

$$\mathbf{X}^{(I\times JK)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} \tag{67}$$

where $\mathbf{A}$ is a 4×2 matrix of random numbers and $\mathbf{B}$ and $\mathbf{C}$ are defined likewise. It is not important how these matrices are generated, as long as they have full column rank. Consider the following alternative two-component PARAFAC models.

1. Using $\underline{\mathbf{X}}$.
2. Using $\underline{\mathbf{X}}$ centered across the first mode and scaled within the combined second and third modes (auto-scaled as a two-way matrix $\mathbf{X}^{(I \times JK)}$.
3. Using $\underline{\mathbf{X}}$ scaled within e.g. mode two.

The fit values of these three models always using two components are given below in percentages of the sum of squares of the preprocessed data.

1. 100.00%.
2. 98.80%.
3. 100.00%.

As can be readily seen, a two-component model is appropriate and should be so even after scaling as in case 3. However, using ordinary two-way scaling methods as in case 2 destroys the multilinear structure of the data and deteriorates the model.

5.3. Simultaneous centering and scaling

The exact same rules for interdependence of preprocessing steps apply for multiway data as for two-way data (Section 4), also with respect to treating missing data, etc. Any preprocessed array may be written in matrix notation using the matricized $I \times JK$ data array $\mathbf{X}$ and the preprocessed array $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{M}_I\mathbf{X}(\mathbf{M}_K \otimes \mathbf{M}_J)^{\mathrm{T}} \tag{68}$$

where e.g. $\mathbf{M}_I$ is an $I \times I$ array holding either the centering or scaling transformation matrix for the first mode or even a combination of such. The exact content of these transformation matrices depends on the type of preprocessing chosen, and in the case of iterative preprocessing, $\mathbf{M}$ may be a product of several matrices [1].

Combined centering and scaling in one operator $\mathbf{M}$ is generally not going to retain all the properties of the two individual operations (see Section 4.3). If centering across the first mode and scaling within the first mode are desired, centering first and scaling afterwards will not lead to an array in which the first-mode vectors are centered, although possible offsets will be eliminated. Scaling first and centering afterwards will not eliminate the original offsets, even though the preprocessed array will have centered first-mode vectors.

Another example based on the guidelines in Section 4 is a situation in which e.g. scaling within the second and third modes is desired together with centering across the first mode. If the preprocessing is performed so that the data are centered after the weights are determined (iteratively) and applied, i.e.

$$\mathbf{Y} = \mathbf{P}^{\perp}[\mathbf{X}(\mathbf{W}_K \otimes \mathbf{W}_J)^{\mathrm{T}}] \tag{69}$$

then the centering operation will destroy the property of e.g. suitable mean square error in the second and third modes (the brackets indicate the proper order of the preprocessing steps). If, on the other hand, the preprocessing is performed as

$$\mathbf{Y} = (\mathbf{P}^{\perp}\mathbf{X})(\mathbf{W}_K \otimes \mathbf{W}_J)^{\mathrm{T}} \tag{70}$$

this is not the case. In this case the data are first centered and then the weights are determined from the centered array rather than from the raw data. Hence the weights in Equation (70) are preferred.

5.4. Summary

Proper single centering of a multiway array can always be expressed as

$$\mathbf{P}^{\perp}\mathbf{X} \tag{71}$$

where $\mathbf{X}$ $(I \times JKL\ldots)$ is a multiway array rearranged to a two-way matrix such that the mode to be centered across is the row mode. Hence $\mathbf{P}^{\perp}$ works on the non-matricized mode (I in this case). All combinations of centering of this form are proper and will maintain a preprocessed array with offsets removed.

Proper scaling of a multiway array can be expressed as

$$\mathbf{W}\mathbf{X} \tag{72}$$

where $\mathbf{X}$ $(I \times JKL\ldots)$ is a multiway array rearranged to a two-way matrix such that the mode to be scaled within is the row mode. Scaling e.g. to unit mean square variation within several modes sequentially will not yield modes within which the variables or samples have unit mean square variation unless iterative determination of the weights is used.

Combinations of centering and scaling are unproblematic only when centering across several modes is desired, or scaling within one mode combined with centering across other modes. Centering must be performed before scaling. All other combinations will only partly fulfil the requirements. For example, centering across a mode followed by scaling within the same mode will not lead to zero-average vectors in the mode, but it will remove any offsets across the mode.

6. CONCLUSION

A number of important features of the common preprocessing steps of centering and scaling have been discussed.

- Centering deals with the structural model; scaling deals with the way in which this model is fitted.
- Centering is part of a two-stage procedure in which offsets are removed first and multilinear terms are estimated in

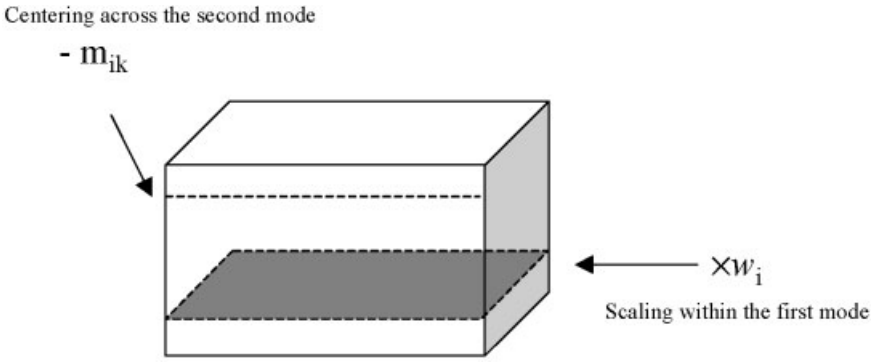

Figure 10. Three-way array. Proper centering must be done across a mode, exemplified here by proper centering across the second mode. From all elements of a specific row (fixed *i* and *k*) the same scalar m_{ik} is subtracted. Proper scaling e.g. within the first mode is performed such that all elements of a specific horizontal slab are multiplied by the same scalar w_i.

the second stage. This is only equivalent to the one-stage procedure of estimating all parameters simultaneously if proper centering, as defined in this paper, is used. Proper centering is shown in Figure 10 (always across one mode at a time).

- For offsets that cannot be removed by using proper centering, a one-stage procedure has to be used. This holds generally for data with missing elements.
- Scaling provides a way to change the objective function by assuming certain weights. Some weight arrangements can be dealt with by scaling followed by ordinary least squares fitting. Only proper scaling is allowed. Proper scaling is shown in Figure 10. For weighting schemes that cannot be dealt with by scaling, weighted least squares algorithms have to be used.
- Incorrect centering or scaling introduces artificial variation. The amount of artificial variation introduced depends on the data and leads to models that are suboptimal to their 'correct' (least squares) counterparts. This is so because the artificial variation has to be modeled additionally.

6.1. Two-way results

- Proper centering can always be written as $\mathbf{P}^{\perp}\mathbf{X}$.
- Several centerings can be performed sequentially.
- Proper scaling can always be expressed as $\mathbf{WX}$.
- Several scalings can be performed sequentially, but will generally need iterations to establish the scaling constants, and this may not converge.
- Unproblematic combinations of centering and scaling can be expressed as $(\mathbf{P}^{\perp}\mathbf{X})\mathbf{W}$. Similar results hold for transposed matrices.

6.2. Three-way results

- Proper centering can always be written as $\mathbf{P}^{\perp}\mathbf{X}$, where $\mathbf{X}$ is the three-way array matricized, so that the mode to be centered across is the first mode.
- Several such single centerings may be performed sequentially across several modes.
- Proper scaling can always be expressed as $\mathbf{WX}$ for a matricized array as above.
- Several scalings can be performed sequentially, but will generally need iterations and may not converge.
- Proper combinations of centering and scaling can be expressed similarly to the two-way case. That is, scaling does not affect centering across other modes, but centering affects scaling within all modes.

The appropriate centering and scaling procedures can most easily be summarized as in Figure 10. Centering must be done by subtracting scalars from individual vectors of the array, while scaling must be performed by multiplying individual slabs.

Acknowledgements

Most algorithms used in this paper are available from the home page of Food Technology at http://www.models.kvl.dk. The first author is grateful for financial support through LMC (Center for Advanced Food Studies), EU (European Union) project NwayQual GRD1-1999-10377 and AQM (Advanced Quality Monitoring) supported by the Danish Ministries of Research and Industry. Anonymous referees are thanked for helpful comments.

APPENDIX I. PROJECTIONS

In this appendix the projection of vectors on other vectors is explained. In Figure 11 the orthogonal projection of an *I*-vector **b** on an *I*-vector **a** is considered. The resulting

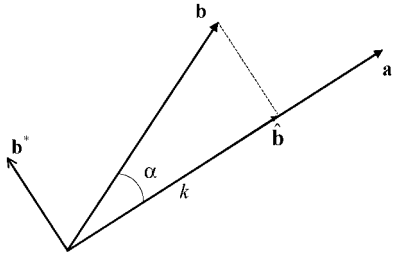

Figure 11. Projection of **b** on **a**.

projection $\hat{\mathbf{b}}$ can be expressed in the original co-ordinate system or in terms of the new basis vector **a**.

I.1. Expression of $\hat{\mathbf{b}}$ in terms of a

The score of $\hat{\mathbf{b}}$ on the new basis vector **a** is k. The following equations hold:

$$\cos\alpha = \frac{k\|\mathbf{a}\|}{\|\mathbf{b}\|}, \qquad \cos\alpha = \frac{(\mathbf{a},\mathbf{b})}{\|\mathbf{a}\|\|\mathbf{b}\|}$$
$$(\mathbf{a},\mathbf{b}) = \mathbf{a}^{\mathrm{T}}\mathbf{b}, \qquad \|\mathbf{a}\| = \sqrt{(\mathbf{a},\mathbf{a})} \tag{73}$$

where the second equation is called the cosine rule for vectors. Hence

$$\frac{k\|\mathbf{a}\|}{\|\mathbf{b}\|} = \frac{(\mathbf{a},\mathbf{b})}{\|\mathbf{a}\|\|\mathbf{b}\|} \Rightarrow k = \frac{(\mathbf{a},\mathbf{b})}{\|\mathbf{a}\|^2} = \frac{(\mathbf{a},\mathbf{b})}{(\mathbf{a},\mathbf{a})} \tag{74}$$

This expression becomes particularly simple when **a** is normalized to length one ($\|\mathbf{a}\| = 1$):

$$k = (\mathbf{a},\mathbf{b}) = \mathbf{a}^{\mathrm{T}}\mathbf{b} \tag{75}$$

Note that expressions like Equation (75) are used in principal component analysis (PCA), where usually the loading vectors **p** are chosen to be of length one and then

$$\mathbf{t} = \mathbf{X}\mathbf{p} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_I^{\mathrm{T}} \end{bmatrix}\mathbf{p} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}}\mathbf{p} \\ \vdots \\ \mathbf{x}_I^{\mathrm{T}}\mathbf{p} \end{bmatrix} \tag{76}$$

where $\mathbf{X}_i^{\mathrm{T}}$ is a row of $\mathbf{X}$ ($I \times J$). Hence the scores of PCA are just orthogonal projections of the rows of $\mathbf{X}$ on **p** in co-ordinates of this **p**.

I.2. Expression of $\hat{\mathbf{b}}$ in terms of the original co-ordinate system

It is also possible to express $\hat{\mathbf{b}}$ in the original co-ordinate system. Referring to Figure 11, it holds that $\hat{\mathbf{b}} = k\mathbf{a}$. Hence

$$\hat{\mathbf{b}} = k\mathbf{a} = \frac{(\mathbf{a},\mathbf{b})}{(\mathbf{a},\mathbf{a})}\mathbf{a} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} = \frac{\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{b}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} = \left(\frac{\mathbf{a}\mathbf{a}^{\mathrm{T}}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\right)\mathbf{b} = \mathbf{P}\mathbf{b} \tag{77}$$

and the matrix **P** ($I \times I$) is special because

$$\mathbf{P}^{\mathrm{T}} = \mathbf{P}$$
$$\mathbf{P}\mathbf{P} = \left(\frac{\mathbf{a}\mathbf{a}^{\mathrm{T}}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\right)\left(\frac{\mathbf{a}\mathbf{a}^{\mathrm{T}}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\right) = \frac{\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{a}\mathbf{a}^{\mathrm{T}}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{a}} = \left(\frac{\mathbf{a}^{\mathrm{T}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\right)\frac{\mathbf{a}\mathbf{a}^{\mathrm{T}}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} = \mathbf{P} \tag{78}$$

which means that **P** is symmetric and idempotent. Then this matrix is an orthogonal projection matrix [31]. It projects orthogonally on the vector **a**. In the case of centering (see Equation (15)), the vector **1** takes the role of **a**.

I.3. Orthogonal projections in general including residuals

A matrix $\mathbf{P} = \mathbf{A}\mathbf{A}^{+}$ (where superscript + indicates the Moore–Penrose inverse) projects orthogonally on the range (column space) of **A** [31]. It can be checked that $\mathbf{a}^{+} = \mathbf{a}^{\mathrm{T}}/(\mathbf{a}^{\mathrm{T}}\mathbf{a})$; hence Equation (78) is a special case of the general orthogonal projection theorem.

It is also interesting to consider the residuals from the orthogonal projection of **b** on **A**; that is, the vector **b***. As $\mathbf{b} = \hat{\mathbf{b}} + \mathbf{b}^{*}$, it holds that

$$\mathbf{b}^{*} = (\mathbf{I} - \mathbf{P})\mathbf{b} = \mathbf{P}^{\perp}\mathbf{b}$$
$$(\mathbf{P}^{\perp})^{\mathrm{T}} = \mathbf{P}^{\perp}$$
$$\mathbf{P}^{\perp}\mathbf{P}^{\perp} = \mathbf{P}^{\perp} \tag{79}$$

and $\mathbf{P}^{\perp}$ is again a symmetric and idempotent matrix, i.e. an orthogonal projection operator. This matrix projects onto the orthogonal complement of **A**, i.e. onto *range*($\mathbf{A}^{\perp}$), where *range*(·) is used to indicate the range (column space) of a matrix or vector. It holds that *range*($\mathbf{A}^{\perp}$) = *nullspace*($\mathbf{A}^{\mathrm{T}}$), where *nullspace*(·) is used to indicate the nullspace of a matrix or vector. For every $\mathbf{x} \in$ *range*($\mathbf{A}^{\perp}$) it holds that $\mathbf{A}^{\mathrm{T}}\mathbf{x} = 0$, hence $\mathbf{x} \in$ *nullspace*($\mathbf{A}^{\mathrm{T}}$), and *vice versa*.

APPENDIX II. FITTING A BILINEAR MODEL PLUS OFFSETS ACROSS ONE MODE EQUALS FITTING A BILINEAR MODEL TO CENTERED DATA

Theorem

Given **X** of size $I \times J$ and the column dimension R of a sought bilinear model. Then

$$min\|\mathbf{X} - (\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{m}^{\mathrm{T}})\|^2 = min\|\mathbf{Y} - \mathbf{C}\mathbf{D}^{\mathrm{T}}\|^2 \tag{80}$$

where **Y** is the original data **X** with the column averages subtracted.

Proof

The proof has been given by Kruskal [32] and Gabriel [33]. Understanding that centering is a projection, it is simple to prove the above theorem. Let the loss function be

$$\|\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{1}\mathbf{m}^{\mathrm{T}}\|^2 \tag{81}$$

and partition it into two orthogonal parts

$$\|\mathbf{P}^{\perp}(\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{1}\mathbf{m}^{\mathrm{T}})\|^2 + \|\mathbf{P}(\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{1}\mathbf{m}^{\mathrm{T}})\|^2 \tag{82}$$

using the Pythagorean fact that the squares of two orthogonal parts equal the square of the total. This equation can be further developed to

$$min\|\mathbf{P}^{\perp}\mathbf{X} - \mathbf{P}^{\perp}\mathbf{A}\mathbf{B}^{\mathrm{T}}\|^2 + min\|\mathbf{P}\mathbf{X} - \mathbf{P}(\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{m}^{\mathrm{T}})\|^2 = min\|\mathbf{P}^{\perp}\mathbf{X} - \mathbf{P}^{\perp}\mathbf{A}\mathbf{B}^{\mathrm{T}}\|^2 \tag{83}$$

because $\|\mathbf{P}\mathbf{X} - \mathbf{P}(\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{m}^{\mathrm{T}})\|^2$ will be zero by setting

$$\mathbf{m}^{\mathrm{T}} = (\mathbf{1}^{\mathrm{T}}/I)(\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}}) \text{ since } \mathbf{1}(\mathbf{1}^{\mathrm{T}}/I) = \mathbf{P} \tag{84}$$

Setting $\mathbf{C} = \mathbf{P}^{\perp}\mathbf{A}$ and $\mathbf{D} = \mathbf{B}$ will therefore provide a solution with exactly the same fit as would be obtained by minimizing the original loss function. The solution may be computed using any bilinear algorithm for fitting a principal component analysis model of **Y**. The scores will automatically be centered, because they are linear combinations of the columns of **X**. If the columns are centered, so are their linear combinations.

REFERENCES

1. Harshman RA, Lundy ME. Data preprocessing and the extended PARAFAC model. In *Research Methods for*

Multimode Data Analysis, Law HG, Snyder Jr CW, Hattie J, McDonald RP (eds). Praeger: New York, 1984; 216–284.
2. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**: 105–122.
3. Kruskal JB. Rank, decomposition, and uniqueness for 3-way and N-way arrays. In *Multiway Data Analysis*, Coppi R, Bolasco S (eds). Elsevier: Amsterdam, 1989; 8–18.
4. Ten Berge JMF. Convergence of PARAFAC preprocessing procedures and the Deming–Stephan method of iterative proportional fitting. In *Multiway Data Analysis*, Coppi R, Bolasco S (eds). Elsevier: Amsterdam, 1989; 53–63.
5. Nørgaard L. Classification and prediction of quality and process parameters of beet sugar and thick juice by fluorescence spectroscopy and chemometrics. *Zuckerindustrie* 1995; **120**: 970–981.
6. Weinberg JR. *A Short History of Medieval Philosophy*. Princeton University Press: Princeton, NJ, 1964; 235–266.
7. Judge GG, Griffifths WE, Carter Hill R, Lütkepohl H, Lee TC. *The Theory and Practice of Econometrics*. Wiley: New York, 1985.
8. Seasholtz MB, Kowalski BR. The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta* 1993; **277**: 165–177.
9. Styan PH. Hadamard products and multivariate statistical analysis. *Linear Algebra Appl.* 1973; **6**: 217–240.
10. Bro R. *Multi-way analysis in the food industry. Models, algorithms, and applications. PhD Thesis*, University of Amsterdam, 1998 (http://www.mli.kvl.dk/staff/foodtech/brothesis.pdf).
11. Tucker LR. *A method for synthesis of factor analysis studies*. Personnel Research Section, Report 984, Department of the Army, 1951.
12. Amrhein M. *Reaction and flow variants/invariants for the analysis of chemical reaction data. PhD Thesis*, Ecole Polytechnique Fédérale de Lausanne, 1998.
13. Amrhein M, Srinivasan B, Bonvin D, Schumacher MM. On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics Intell. Lab. Syst.* 1996; **33**: 17–33.
14. Grung B, Manne R. Missing values in principal component analysis. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 125–139.
15. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
16. Cornelius PL, Seyedsadr M, Crossa J. Using the shifted multiplicative model to search for 'separability' in crop cultivar trials. *Theor. Appl. Genet.* 1992; **84**: 161–172.
17. Van Eeuwijk FA. *Between and beyond additivity and non-additivity; the statistical modelling of genotype by environment interaction in plant breeding. PhD Thesis*, University of Wageningen, 1996.
18. Westerhuis JA, Kourti T, MacGregor JF. Comparing alternative approaches for multivariate statistical analysis of batch process data. *J. Chemometrics* 1999; **13**: 397–413.
19. Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 1994; **5**: 111–126.
20. Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* 1989; **43**: 772–777.
21. Guo Q, Wu W, Massart DL. The robust normal variate transform for pattern recognition with near-infrared data. *Anal. Chim. Acta* 1999; **382**: 87–103.
22. Helland IS, Næs T, Isaksson T. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics Intell. Lab. Syst.* 1995; **29**: 233–241.
23. Bro R, Sidiropoulos ND, Smilde AK. Maximum likelihood fitting using simple least squares algorithms. *J. Chemometrics* 2002; **16**: 387–400.
24. Wentzell PD, Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemometrics Intell. Lab. Syst.* 1999; **45**: 65–85.
25. Kiers HAL. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika* 1997; **62**: 251–266.
26. Paatero P. A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 223–242.
27. Andrews DT, Wentzell PD. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. *Anal. Chim. Acta* 1997; **350**: 341–352.
28. Wentzell PD, Andrews DT, Hamilton DC, Faber NM, Kowalski BR. Maximum likelihood principal component analysis. *J. Chemometrics* 1997; **11**: 339–366.
29. Ten Berge JMF, Kiers HAL. Convergence properties of an iterative procedure of ipsatizing and standardizing a data matrix, with application to PARAFAC/CANDECOMP preprocessing. *Psychometrika* 1989; **54**: 231–235.
30. Kruskal JB. Multilinear methods. *Proc. Symp. Appl. Math.* 1983; **28**: 75–104.
31. Schott JR. *Matrix Analysis for Statistics*. Wiley: New York, 1997.
32. Kruskal JB. *Some least squares theorems for matrices and N-way arrays. Manuscript*, Bell Laboratories, Murray Hill, NJ, 1977.
33. Gabriel KR. Least squares approximation of matrices by additive and multiplicative models. *J. R. Statist. Soc. B* 1978; **40**: 186–196.

