JOURNAL OF CHEMOMETRICS
J. Chemometrics 2002; **16**: 387–400
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.734

Maximum likelihood fitting using ordinary least squares algorithms†

Rasmus Bro[1]*, Nicholaos D. Sidiropoulos[2] and Age K. Smilde[3]

[1]Chemometrics Group, Food Technology, Department of Dairy and Food Science, Royal Veterinary and Agricultural University, Frederiksberg, Denmark
[2]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[3]Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Amsterdam, The Netherlands

Received 10 August 2001; Revised 27 March 2002; Accepted 4 April 2002

In this paper a general algorithm is provided for maximum likelihood fitting of deterministic models subject to Gaussian-distributed residual variation (including any type of non-singular covariance). By deterministic models is meant models in which no distributional assumptions are valid (or applied) on the parameters. The algorithm may also more generally be used for weighted least squares (WLS) fitting in situations where either distributional assumptions are not available or other than statistical assumptions guide the choice of loss function. The algorithm to solve the associated problem is called MILES (Maximum likelihood via Iterative Least squares EStimation). It is shown that the sought parameters can be estimated using simple least squares (LS) algorithms in an iterative fashion. The algorithm is based on iterative majorization and extends earlier work for WLS fitting of models with heteroscedastic uncorrelated residual variation. The algorithm is shown to include several current algorithms as special cases. For example, maximum likelihood principal component analysis models with and without offsets can be easily fitted with MILES. The MILES algorithm is simple and can be implemented as an outer loop in any least squares algorithm, e.g. for analysis of variance, regression, response surface modeling, etc. Several examples are provided on the use of MILES.

KEYWORDS: iterative majorization; PARAFAC; PCA; weighted least squares; measurement error

1. INTRODUCTION

In order to set the stage for the maximum likelihood method developed here, it is necessary to first discuss the least squares and weighted least squares approaches for modeling. Often, least squares fitting is used for estimating parameters. Least squares fitting is especially useful for fitting parameters when the residual variation is homoscedastic, independent and Gaussian. In this case, least squares fitting will yield maximum likelihood estimates. When the residual variation is no longer homoscedastic, the different magnitudes of different errors can be handled by using *weighted* least squares fitting rather than least squares fitting. The weight attached to a certain residual reflects the inverse of the known, expected or estimated uncertainty of the corresponding data element, i.e. the inverse uncertainty variance of that particular residual. However, in some situations the errors are not independent, and simple weighted least squares is no longer optimal. In order to handle correlated errors, more complicated fitting procedures are generally required, and often the algorithms for such fitting procedures are either complicated to implement or cumbersome to use. In this paper a general fitting procedure will be devised that can handle all the above situations for fitting any model for which a least squares fitting procedure exists.

A short generic description of the algorithm is given at this stage to help understand the subsequent derivations. A table of uncertainty information $\mathbf{W}$ is available which defines weights for each individual element of the input data $\mathbf{X}$. This weight matrix can hold information about the covariance of the uncertainty between different elements. The purpose of MILES is to allow for this information to be used in the modeling of the data. The data can have any suitable structure, e.g. two-way (as in $\mathbf{X}$) or a vector (as in $\mathbf{x}$), a three-way array (as in $\underline{X}$) or any other structure. MILES is an iterative procedure which alternates between a modification of the original input data $\mathbf{X}$ by so-called *majorization* and a *conventional least squares modeling* of the modified data. The latter takes place in conventional space, while the former takes place in a one-way representation of the data. This one-

*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: rb@kvl.dk
†Paper presented at the 7th Scandinavian Symposium on Chemometrics, Copenhagen, Denmark, 19–23 August 2001.
Contract/grant sponsor: LMC (Centre for Advanced Food Studies).
Contract/grant sponsor: EU (European Union); Contract/grant number: GRD1-1999-10337.
Contract/grant sponsor: NSF (National Science Foundation); Contract/grant number: 0096164; Contract/grant number: 0096165.

way representation is obtained by stringing out the data, e.g. $\mathbf{X}$, into a long column vector $\mathbf{x}$ which simply holds all the columns in $\mathbf{X}$. Thus there is a one-to-one mapping between $\mathbf{X}$ and $\mathbf{x}$ by a simple rearrangement. In each step of MILES the data are first modified using the current interim model, called $\mathbf{M}$ or $\mathbf{m}$ depending on whether the representation is in conventional or vectorized space. The modified data called $\mathbf{q}$ are rearranged to the original format $\mathbf{Q}$, and in the second step the model is simply updated by performing an ordinary least squares fitting of the model to $\mathbf{Q}$ (instead of $\mathbf{X}$). Upon convergence a model is obtained which takes the weights into account.

The general models dealt with in this paper are described along with the optimization problem to be solved. Let $\mathbf{x}$ be an $I \times 1$ vector holding a data set for which a model is sought. Having $\mathbf{x}$ organized in a vector is not a restriction of the type of data that can be modeled. This vector may represent a matrix, a three-way array, a multi-way array or any other structure of data elements rearranged appropriately into a vector. For a $P \times J$ matrix $\mathbf{Z}$ this rearrangement can be achieved by vectorizing the matrix such that $\mathbf{x} = \mathrm{vec}\mathbf{Z} = [\mathbf{z}_1^{\mathrm{T}}\ \mathbf{z}_2^{\mathrm{T}} \ldots\ \mathbf{z}_J^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{z}_j$ is a P-vector holding the jth column of $\mathbf{Z}$. For a three-way $P \times J \times K$ array $\underline{Z}$, vectorization can be done by setting $\mathbf{x} = \mathrm{vec}\underline{Z} = [\mathrm{vec}\mathbf{Z}_1^{\mathrm{T}}\ \mathrm{vec}\mathbf{Z}_2^{\mathrm{T}} \ldots\ \mathrm{vec}\mathbf{Z}_K^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{Z}_k$ is a $P \times J$ matrix with typical elements z_{pjk}, $p = 1, \ldots, P$, $j = 1, \ldots, J$.

A model $\mathbf{m}$ ($I \times 1$) with a certain structure and certain constraints is sought for the data. Hence

$$\mathbf{x} = \mathbf{m} + \mathbf{e} \tag{1}$$

where $\mathbf{e}$ ($I \times 1$) holds the residual variation not explained by the model. The model $\mathbf{m}$ is defined as belonging to a certain set, i.e. $\mathbf{m} \in \Upsilon$, where the set Υ defines the structure and constraints (e.g. PCA). In fitting the model in a least squares sense, $\mathbf{m}$ is found as the argument minimizing

$$\sigma_{\mathrm{LS}}(\mathbf{m}|\mathbf{x}) = \|\mathbf{x} - \mathbf{m}\|^2 \tag{2}$$

over $\mathbf{m} \in \Upsilon$, where $\|\mathbf{g}\|^2$ denotes the squared Euclidean/Frobenius norm of $\mathbf{g}$, i.e. the sum of the squared elements of $\mathbf{g}$. The expression $\sigma_{\mathrm{LS}}(\mathbf{m} \parallel \mathbf{x})$ indicates that the loss function σ_{LS} is a function of both $\mathbf{m}$ and $\mathbf{x}$ and that $\mathbf{x}$ is considered fixed; hence only the model $\mathbf{m}$ is optimized. This follows the notation used by Kiers in a related paper [1]. For the model $\mathbf{m}$, any structure and/or constraints can apply.

For example, in principal component analysis, $\mathbf{m} = \mathrm{vec}\mathbf{M} = \mathrm{vec}(\mathbf{A}\mathbf{B}^{\mathrm{T}})$, with both $\mathbf{A}$ and $\mathbf{B}$ having orthogonal columns. Centering can also be easily included. Fitting e.g. a PCA model to centered data is a two-step procedure, where offsets are first removed (centering operation) and the PCA model is subsequently fitted to the centered data. However, several authors have proven that this two-step procedure is equivalent to fitting a certain well-defined model to the raw data in a *least squares sense* [2–4]. This model can be written $\mathbf{M} = \mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{1}\mathbf{n}^{\mathrm{T}}$, where $\mathbf{n}$ is a J-vector holding the offsets (the average of the jth column in the case where these are estimated by centering). Thus, instead of first centering and subsequently fitting a model, it is equally valid to specify the whole model as one overall least squares problem. Hence least squares models involving centering are immediately covered by the methodology developed in this paper.

Yet another model could be a three-way PARAFAC model [5], in which case $\mathbf{M} = [\mathbf{A}\mathbf{D}_1\mathbf{B}^{\mathrm{T}}\ \mathbf{A}\mathbf{D}_2\mathbf{B}^{\mathrm{T}} \ldots\ \mathbf{A}\mathbf{D}_K\mathbf{B}^{\mathrm{T}}]$, with $\mathbf{D}_k$ ($k = 1,\ldots,K$) being diagonal. As yet another alternative, the model could be a multivariate linear regression model, which can be formulated as $\mathbf{M} = \mathbf{Q}\mathbf{B}$, where $\mathbf{Q}$ is a given fixed $I \times P$ matrix holding the independent variables, $\mathbf{M}$ ($I \times R$) holds the model of the R dependent variables and $\mathbf{B}$ ($P \times R$) holds the regression coefficients. In short, any problem which has a least squares formulation can be expressed according to Equation (2). For many special types of models there are well-established algorithms for fitting the above least squares model. This paper deals with incorporating knowledge of the error covariance structure or *a priori* problem-specific information in finding the solution to Equation (1), taking advantage of the availability of an algorithm for fitting the least squares model.

A way to pose the problem considered is the following. The problem is to be able to fit deterministic models (i.e. where parameters have no distributional assumptions attached) but subject to any type of non-singular covariance of the residual Gaussian noise. A different problem that leads to the same loss function and hence solution is the problem of fitting a model in a weighted least squares (WLS) sense, i.e. to minimize the loss $\mathbf{e}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{e}$, where $\mathbf{e}$ is the residual vector and $\mathbf{W}$ is a non-singular weight matrix. In the following the derivation of the algorithm will be presented as the problem of maximum likelihood fitting, but, as stated, weighted least squares fitting can also be sensible on the basis of non-distributional *a priori* information, even if it does not admit a maximum likelihood interpretation.

After deriving the MILES algorithm, several applications are discussed. The applications will focus on well-known chemometric models such as principal component analysis and PARAFAC.

2. THEORY

2.1. Problem formulation

It is assumed that the residual vector $\mathbf{e}$ ($I \times 1$) in the model $\mathbf{x} = \mathbf{m} + \mathbf{e}$ is zero-mean Gaussian, with known covariance matrix

$$\mathrm{cov}(\mathbf{e}) = \mathbf{\Delta} = E(\mathbf{e}\mathbf{e}^{\mathrm{T}}) \tag{3}$$

where $\mathbf{\Delta}$ is a matrix of size $I \times I$ assumed to be of full rank and $\mathbf{e}^{\mathrm{T}}$ is the transpose of $\mathbf{e}$. If $\mathbf{e}$ does not have zero mean, but the mean is known or can be estimated, this mean can simply be subtracted from the observations. The vector $\mathbf{m}$ is a deterministic unknown subject to structural constraints. Specifically, the only thing known about $\mathbf{m}$ is that it belongs to the set Υ, but no distribution of the outcomes of $\mathbf{m}$ over Υ is known or can be assumed. We will develop a deterministic likelihood algorithm for fitting this model.

Estimating parameters in a model using least squares algorithms is appropriate for situations in which the errors $\mathbf{e}$ in Equation (1) are i.i.d. Gaussian. By appropriate is meant that the maximum likelihood solution is found under the given premises. However, for other types of distribution this is not the case. If the residuals are independently Gaussian distributed but with different variances, the maximum likelihood principle leads to fitting the model in a weighted

least squares sense according to the criterion

$$\sigma_{\mathrm{WLS}}(\mathbf{m}|\mathbf{x},\mathbf{W}) = \|\mathbf{W}(\mathbf{x}-\mathbf{m})\|^2 \quad (4)$$

where $\mathbf{W}$ is an $I \times I$ *diagonal* matrix holding the weights and $\|\mathbf{G}\|^2$ is the squared Frobenius norm of $\mathbf{G}$. The ith diagonal element signifies the uncertainty of the ith element of $\mathbf{x}$ and is equal to the inverse of the standard deviation of the corresponding residual element. In e.g. bilinear decomposition such as principal component analysis (PCA) it is common to fit models using this criterion in the special situation where the elements in $\mathbf{W}$ corresponding to one specific column of the data matrix have the same value. The parameters of the model can then be found by fitting an ordinary least squares model to the data appropriately preprocessed by scaling each column [2,6]. For fitting PCA models in the case of general but diagonal $\mathbf{W}$, more advanced algorithms have to be used [7,8]. An algorithm for fitting any structural model according to this criterion has been devised by Kiers [1] and, in fact, the work presented here can be seen as a natural extension of his work. A different algorithm has been proposed specifically for PCA by Wentzell *et al.* [9], who also extended their model to the more complicated problem of having correlated errors in PCA.

If there is covariance between the different elements of $\mathbf{e}$, the assumption of independence of the distributions of individual residuals is not valid. Then the loss function in Equation (4) with only diagonal $\mathbf{W}$ no longer yields the maximum likelihood solution. When $\mathbf{e}$ is zero-mean multivariate Gaussian, the covariance of its distribution is

$$\mathrm{cov}(\mathbf{e}) = \boldsymbol{\Delta} \quad (5)$$

which has non-zero off-diagonal elements in general and is assumed to be known. This covariance structure incorporates the assumed error distribution for both σ_{LS} and σ_{WLS} as extreme cases.

The relevant part of the log-likelihood function of estimating the model parameters $\mathbf{m}$ is (Reference [10], p. 181)

$$(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\Delta^{-1}(\mathbf{x}-\mathbf{m}) \quad (6)$$

and by taking $\mathbf{W} = \Delta^{-1/2}$, Equation (6) can be rewritten as

$$(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}) \quad \text{or} \quad [\mathbf{W}(\mathbf{x}-\mathbf{m})]^{\mathrm{T}}[\mathbf{W}(\mathbf{x}-\mathbf{m})] \quad (7)$$

Maximizing the likelihood requires minimizing Equation (7). Hence the model can be found by minimizing the loss function (see e.g. Reference [11])

$$\sigma_{\mathrm{ML}}(\mathbf{m}|\mathbf{x},\mathbf{W}) = \|\mathbf{W}(\mathbf{x}-\mathbf{m})\|^2 \quad (8)$$

The matrix $\mathbf{W}$ is of size $I \times I$ and holds the weights. In the case of the error distribution of Equation (5) it holds that $\mathbf{W} = \Delta^{-1/2}$, assuming full rank of Δ, i.e. $\mathbf{W}$ is the symmetric square root of the inverse of Δ. Thus, if the eigendecomposition of Δ^{-1} is $\Delta^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, with $\mathbf{D}$ diagonal and $\mathbf{U}$ an orthogonal matrix holding the eigenvectors, then $\mathbf{W} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^{\mathrm{T}}$.

As stated in Section 1, the algorithm developed here can also be applied for weighted least squares fitting in situations where the weights do not arise (solely) from statistical considerations. The basis for this broader viewpoint is the loss function in Equation (8). Rather than deriving the weights in $\mathbf{W}$ from statistical considerations, it is possible to introduce any other suitable set of weights based e.g. on *a priori* problem-specific information. Minimizing the loss function in this case will lead to weighted least squares estimates that do not necessarily have any maximum likelihood properties. An example of this is provided in Section 3.

The loss function can be rewritten as

$$\begin{aligned}\sigma_{\mathrm{ML}}(\mathbf{m} \mid \mathbf{x},\mathbf{W}) &= \|\mathbf{W}(\mathbf{x}-\mathbf{m})\|^2 \\ &= (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m})\end{aligned} \quad (9)$$

It is the objective in this paper to derive an algorithm for finding $\mathbf{m}$ with its associated structure and constraints that will minimize σ_{ML}. This is pursued in the next subsection.

For some models it is very simple to devise a maximum likelihood algorithm from a least squares algorithm. For example, consider the multiple linear regression problem

$$\begin{aligned}\sigma_{\mathrm{ML}}(\mathbf{b} \mid \mathbf{x},\mathbf{Z},\mathbf{W}) &= \|\mathbf{W}(\mathbf{x}-\mathbf{Z}\mathbf{b})\|^2 \\ &= (\mathbf{x}-\mathbf{Z}\mathbf{b})^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{Z}\mathbf{b})\end{aligned} \quad (10)$$

It is easy to see that the solution can be found as the solution to the ordinary least squares problem

$$\begin{aligned}\sigma_{\mathrm{ML}}(\mathbf{b} \mid \mathbf{x},\mathbf{Z},\mathbf{W}) &= \|\mathbf{W}(\mathbf{x}-\mathbf{Z}\mathbf{b})\|^2 \\ &= (\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{Z}\mathbf{b})^{\mathrm{T}}(\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{Z}\mathbf{b})\end{aligned} \quad (11)$$

by regressing $\mathbf{W}\mathbf{x}$ onto $\mathbf{W}\mathbf{Z}$. This is called generalized least squares and is standard in regression theory (Reference [12], p. 266). It is also called whitening in filtering theory [13]. Also, consider a bilinear PCA model of an $I \times J$ matrix $\mathbf{X}$. If the errors are *independent* across rows and identically distributed within each row, the weight matrix $\mathbf{W}$ is a block-diagonal matrix with blocks of $J \times J$ matrices $\mathbf{V}$ that are all identical. It can be shown [6] that in such a case the problem is equivalent to minimizing the loss function

$$\begin{aligned}\sigma_{\mathrm{ML}}(\mathbf{T},\mathbf{P} \mid \mathbf{X},\mathbf{V}) &= \|(\mathbf{X}-\mathbf{T}\mathbf{P}^{\mathrm{T}})\mathbf{V}\|^2 \\ &= \mathrm{vec}[(\mathbf{X}-\mathbf{T}\mathbf{P}^{\mathrm{T}})\mathbf{V}]^{\mathrm{T}}\mathrm{vec}[(\mathbf{X}-\mathbf{T}\mathbf{P}^{\mathrm{T}})\mathbf{V}]\end{aligned} \quad (12)$$

The solution can be obtained by fitting a PCA model to the transformed data according to

$$\sigma_{\mathrm{ML}}(\mathbf{T},\mathbf{P} \mid \mathbf{X},\mathbf{V}) = \mathrm{vec}(\mathbf{X}\mathbf{V}-\mathbf{T}\mathbf{P}^{\mathrm{T}}\mathbf{V})^{\mathrm{T}}\mathrm{vec}(\mathbf{X}\mathbf{V}-\mathbf{T}\mathbf{P}^{\mathrm{T}}\mathbf{V}) \quad (13)$$

The parameters are thus obtained as $\mathbf{T}$ and $\mathbf{P}^{\mathrm{T}}\mathbf{V}$ when fitted to $\mathbf{X}\mathbf{V}$, and $\mathbf{P}$ can be found as $(\mathbf{P}^{\mathrm{T}}\mathbf{V}\mathbf{V}^{-1})^{\mathrm{T}}$. Because $(\mathbf{P}^{\mathrm{T}}\mathbf{V}\mathbf{V}^{-1})^{\mathrm{T}}$ is generally not orthogonal, the parameters have to be appropriately orthogonalized in order to get the standard PCA solution. This can most easily be obtained from a singular value decomposition of $\mathbf{T}\mathbf{P}^{\mathrm{T}}\mathbf{V}\mathbf{V}^{-1}$. Such preprocessing can only be used when the errors are independent across all but one mode. For general $\mathbf{W}$ of size $I \times I$, even PCA models cannot be fitted using least squares algorithms on preprocessed data. For e.g. multi-way models such as PARAFAC [5] and Tucker3 [14] the problems become even more pronounced.

2.2. Deriving an algorithm

An algorithm for minimizing σ_{ML} ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$) can be developed using iterative majorization. The theory of majorization has been described in the literature [15–18], but the principle as it pertains to minimizing a loss function is briefly reviewed next.

The loss function can be minimized by iteratively improving any current estimate of the model. Let such a current estimate be $\mathbf{m}_c$, where c is the iteration number. An update is sought such that $\sigma_{ML}(\mathbf{m}_{c+1} \mid \mathbf{x},\mathbf{W}) < \sigma_{ML}(\mathbf{m}_c \mid \mathbf{x},\mathbf{W})$. Improving the estimate continuously will lead to convergence, because the loss is bounded below by zero. For a complicated problem, however, improving the estimate can be difficult. If no closed-form solution exists, gradient-based methods may be used. Majorization, though, provides another alternative route for improving a current estimate. Instead of optimizing the loss function directly, a new loss function is derived which is called the *majorizing function*. This is a function of $\mathbf{x}$, $\mathbf{W}$ *and* the current estimate $\mathbf{m}_c$. The majorizing function is denoted by σ_{maj} ($\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_c$), indicating that the aim is to find an estimate $\mathbf{m}$ but now also as a function of the current estimate $\mathbf{m}_c$. The majorizing function is constructed in such a way that (i) it is easy (or easier) to minimize and (ii) an improvement of σ_{maj} ($\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_c$) will also lead to an improvement of σ_{ML} ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$). How this is achieved is described below and a pictorial description of majorization can also be seen in Figure 1.

The majorization function should satisfy certain properties to lend itself to iterative minimization of the original loss function. It must hold that the value of the majorizing function is never smaller than that of the ML loss function to ensure monotonic convergence; that is, σ_{maj} ($\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_{c+1}$) $\geq \sigma_{ML}$ ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$), $\forall \mathbf{m} \in \Upsilon$. This requirement is the reason for the name majorizing function. In order to obtain a reasonable convergence rate, it is often appropriate, though, that σ_{maj} ($\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_{c+1}$) is close to σ_{ML} ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$) at least in the neighborhood of $\mathbf{m}_c$. Further, it must hold that the two loss functions are identical at the *supporting point*, which is the current estimate $\mathbf{m}_c$, hence that σ_{maj} ($\mathbf{m}_c \mid \mathbf{x},\mathbf{W},\mathbf{m}_c$) = σ_{ML} ($\mathbf{m}_c \mid \mathbf{x},\mathbf{W}$). Note that this supporting point will not be the point corresponding to the minimum of either σ_{maj} ($\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_{c+1}$) or σ_{ML} ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$) unless at convergence.

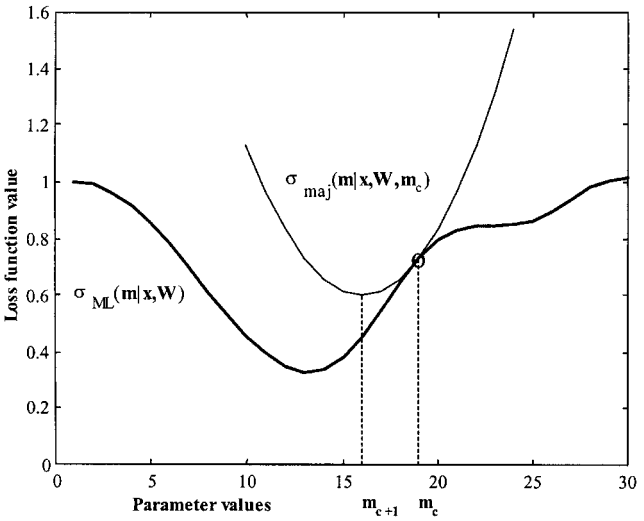

Figure 1. The principle behind majorization illustrated with a one-parameter model. The loss function σ_{ML} is to be minimized as a function of $\mathbf{m}$. The current estimate of $\mathbf{m}$ is $\mathbf{m}_c$ (abscissa) with a corresponding loss (ordinate). This loss is improved by minimizing a majorizing function σ_{maj}.

A majorizing function for σ_{ML} ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$) will now be developed fulfilling the above criteria. This majorizing function is closely related to the one provided by Kiers [1] and uses the function suggested by Heiser [19] in another context. The derivation here follows closely that of these two papers.

As shown in Equation (9), the loss function σ_{ML} ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$) can be written as

$$\begin{aligned}\sigma_{ML}(\mathbf{m} \mid \mathbf{x},\mathbf{W}) &= \|\mathbf{W}(\mathbf{x}-\mathbf{m})\|^2 \\ &= (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}) \end{aligned} \tag{14}$$

The sought model $\mathbf{m}$ can be written as $\mathbf{m} = \mathbf{m}_c + (\mathbf{m} - \mathbf{m}_c)$, where $\mathbf{m}_c$ is the current estimate of $\mathbf{m}$. Thus Equation (14) can be formulated as

$$\begin{aligned}\sigma_{ML}(\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_c) &= [(\mathbf{x}-\mathbf{m}_c) \\ &\quad -(\mathbf{m}-\mathbf{m}_c)]^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}[(\mathbf{x}-\mathbf{m}_c)-(\mathbf{m}-\mathbf{m}_c)] \\ &= (\mathbf{x}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}_c) \\ &\quad +(\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{m}-\mathbf{m}_c) \\ &\quad -2(\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}_c) \end{aligned} \tag{15}$$

Note that the first term $\alpha = (\mathbf{x}-\mathbf{m}_c)^{\mathrm{T}}\,\mathbf{W}^{\mathrm{T}}\mathbf{W}\,(\mathbf{x}-\mathbf{m}_c)$ is a constant because $\mathbf{x}$ and $\mathbf{m}_c$ are known and fixed, i.e.

$$\begin{aligned}\sigma_{ML}(\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_c) &= \alpha + (\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{m}-\mathbf{m}_c) \\ &\quad -2(\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}_c) \end{aligned} \tag{16}$$

The third term is linear in $\mathbf{m}$, while the second term is the difficult part because it is quadratic in $\mathbf{m}$ and because of the presence of the matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}$. As shown by Heiser [19], modifying this second term can provide a majorizing function well suited for the purpose here. Let β be the largest eigenvalue of $\mathbf{W}^{\mathrm{T}}\mathbf{W}$. It then holds that

$$\beta = \max_{\mathbf{s}}\left(\frac{\mathbf{s}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{s}}{\mathbf{s}^{\mathrm{T}}\mathbf{s}}\right) \tag{17}$$

for vectors $\mathbf{s}$ of appropriate size. Thus for any $\mathbf{s}$ it holds that

$$\beta\mathbf{s}^{\mathrm{T}}\mathbf{s} \geqslant \mathbf{s}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{s} \tag{18}$$

Therefore it holds that

$$\beta\,(\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}(\mathbf{m}-\mathbf{m}_c) \geqslant (\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{m}-\mathbf{m}_c) \tag{19}$$

and from this relation a majorizing function can be defined from Equation (16) as

$$\begin{aligned}\sigma_{maj}(\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_c) &= \alpha + \beta\,(\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}(\mathbf{m}-\mathbf{m}_c) \\ &\quad -2\,(\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}_c) \end{aligned} \tag{20}$$

It follows that σ_{maj} ($\mathbf{m} \mid \mathbf{x},\mathbf{W},\mathbf{m}_c$) $\geq \sigma_{ML}$ ($\mathbf{m} \mid \mathbf{x},\mathbf{W}$) for all $\mathbf{m}$. Setting $\mathbf{m} = \mathbf{m}_c$ leads to σ_{maj} ($\mathbf{m}_c \mid \mathbf{x},\mathbf{W},\mathbf{m}_c$) = σ_{ML} ($\mathbf{m}_c \mid \mathbf{x},\mathbf{W}$) = α. Thus the requirements of a majorizing function are satisfied. In order to improve the current estimate of $\mathbf{m}$ with

respect to the loss function in Equation (16), it suffices to improve $\mathbf{m}$ with respect to Equation (20). This provides a tremendous simplification, because, as will be shown, finding the minimum of Equation (20) corresponds to solving a certain ordinary least squares problem.

Defining the constant vector $\mathbf{q} = \mathbf{m}_c + (1/\beta)\,\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{m}_c)$, the majorizing function can be written as

$$\sigma_{\mathrm{maj}}(\mathbf{m} \mid \mathbf{x}, \mathbf{W}, \mathbf{m}_c) = \delta + \beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m} - 2\ \beta\ \mathbf{m}^{\mathrm{T}}\mathbf{q} \qquad (21)$$

The proof of this is provided in the Appendix. In finding the minimum of this loss function, we can ignore the constant δ and β, and the argument $\mathbf{m}$ that minimizes this function is thus also the solution to

$$\min\ [(\mathbf{m} - \mathbf{q})^{\mathrm{T}}(\mathbf{m} - \mathbf{q})] = \min\ (\|\mathbf{m} - \mathbf{q}\|^2) \qquad (22)$$

It therefore holds that, in order to improve an estimated model $\mathbf{m}_c$ with respect to $\sigma_{\mathrm{ML}}(\mathbf{m}_c \mid \mathbf{x},\mathbf{W})$, it suffices to find the updated model $\mathbf{m}_{c+1}$ that minimizes $\sigma_{\mathrm{LS}}(\mathbf{m} \mid \mathbf{q}) = \|\mathbf{m} - \mathbf{q}\|^2$ for $\mathbf{m} \in \Upsilon$ and with $\mathbf{q} = \mathbf{m}_c + (1/\beta)\,\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{m}_c)$. Thus a least squares model fitted to the transformed data $\mathbf{q}$ will provide the necessary update. The monotonic convergence for an algorithm based on iterative majorization, such as the one derived here, follows, as shown by Kiers [1], from the fact that

$$\begin{aligned}\sigma_{\mathrm{ML}}(\mathbf{m}_{c+1} \mid \mathbf{x}, \mathbf{W}) &\leqslant \sigma_{\mathrm{maj}}(\mathbf{m}_{c+1} \mid \mathbf{x}, \mathbf{W}, \mathbf{m}_c) \\ &\leqslant \sigma_{\mathrm{maj}}(\mathbf{m}_c \mid \mathbf{x}, \mathbf{W}, \mathbf{m}_c) = \sigma_{\mathrm{ML}}(\mathbf{m}_c \mid \mathbf{x}, \mathbf{W})\end{aligned}$$

A convergent algorithm for finding the deterministic maximum likelihood model follows immediately.

Algorithm MILES

1. Set counter $c := 0$; initialize model $\mathbf{m}_c$, e.g. using the least squares model.
2. $\mathbf{q} = \mathbf{m}_c + (1/\beta)\ \mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{m}_c)$, where $\beta = \max[\mathrm{eigenvalue}(\mathbf{W}^{\mathrm{T}}\mathbf{W})]$.
3. $\mathbf{m}_{c+1} = \arg\min_{\mathbf{m} \in \Upsilon} (\|\mathbf{m} - \mathbf{q}\|^2)$
4. $c := c + 1$; go to step 2 until $\|\mathbf{m}_c - \mathbf{m}_{c-1}\|^2 / \|\mathbf{m}_{c-1}\|^2 \leq \varepsilon$, where ε is a pre-specified small number (e.g. 10^{-6}).

In the case of iterative least squares algorithms such as PARAFAC based on alternating least squares, step 3 can often be improved. Instead of finding the least squares update, it is sufficient to find an update that improves the fit. For a slowly converging least squares algorithm it may for example suffice to do 10 iterations or perhaps only one iteration and then subsequently update $\mathbf{q}$. In order to ensure reasonable convergence, it is useful to add a simple line search (on $\mathbf{q}$), and for minimizing the risk of encountering local minima, it is useful to redo the analysis a few times from different starting points. In step 4 the convergence is determined in terms of the relative change in the vectorized model. The algorithm is guaranteed to converge in (WLS/ML) fit, but convergence in fit does not in general imply convergence of model parameters or the vectorized model *per se*. In practical applications of MILES, convergence will likely be determined via relative change tests on the vectorized model, for the sake of computational simplicity. It is for this reason that this choice of convergence criterion is incorporated in MILES.

Computational complexity is an important aspect of any algorithm. MILES is closely related to the principle of alternating least squares (ALS) and, as for ALS, MILES can at most achieve linear convergence. The complexity of MILES will equal the complexity of the least squares algorithm times the number of MILES 'correction' iterations. The number of MILES iterations will of course vary depending on the problem, but is typically fairly high (>100). The key benefit that MILES brings to the table is ease of implementing maximum likelihood fitting of models for which least squares fitting is already available, and where developing a dedicated maximum likelihood fitting routine would be cumbersome and time-consuming. Hence it is not expected that MILES will be faster than dedicated maximum likelihood fitting routines, when these become available, although MILES often matches the speed of existing maximum likelihood fitting routines for some well-known models. To reiterate, MILES offers programming convenience and the ability to test the potential benefits of maximum likelihood fitting before actually developing a dedicated algorithm, which might be faster than MILES but at a significant development cost.

If MILES converges to the global optimum (i.e. it is a maximum likelihood solution), then the good properties of maximum likelihood carry over to the solution obtained through MILES. That is, for large total sample sizes (*PJK* in PARAFAC with a $P \times J \times K$ array) relative to F (rank in PARAFAC), maximum likelihood is asymptotically statistically efficient, i.e. it comes close to achieving the Cramer–Rao lower bound on the variance of all unbiased estimators of the model parameters. This is also true for high signal-to-noise ratios [20].

3. APPLICATION OF MILES

In the following, two brief examples and a more detailed example will be provided showing the usefulness of MILES and comparing MILES with other similar algorithms. The first example is devoted to comparing MILES with maximum likelihood PCA on a published data set. Few practical details will be added to the application to these data, because the data set and associated results have already been excellently described in the original papers. The example mainly serves to verify on known data that MILES provides the same results as those obtained with the already existing algorithm ML-PCA, as well as to illustrate that incorporation of centering is simple in the MILES approach. The second example shows the usefulness of MILES in maximum likelihood PARAFAC modeling in signal processing with correlated errors, while the final more detailed example shows that MILES can also be used as a more general weighted least squares approach for handling measurement artifacts in a situation where maximum likelihood fitting does not apply.

3.1. Spectroscopy—PCA

The first data set was generated and treated by Wentzell and Lohnes [21] and arose from a designed experiment with three-component mixtures of nitrates of Co(II), Cr(III) and Ni(II). A three-level, three-factor calibration design was used

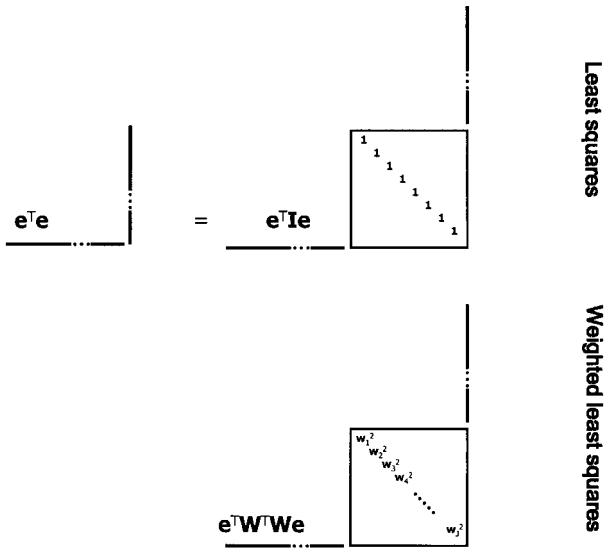

Figure 2. The loss function in least squares fitting and weighted least squares fitting. The residuals are given in the vector **e**, which is equal to vec(**X-M**), where **M** is the PCA model (including possible offsets). Thus, in weighted least squares fitting, one specific weight is attached to each residual and hence to each element in **X**. ML-PCA and MILES-PCA as given below also handle non-diagonal **W**, but this is not introduced in the current application.

in which 1, 3 or 5 ml aliquots of stock solutions of the three nitrates were combined and diluted to 25 ml with 4% nitric acid. For technical reasons, one sample was not measured, so the data set consists of 26 rather than 27 samples. In the original publication, two samples were left out of the data set because their standard deviations were judged to be extreme. It was noted that inclusion did not change the results significantly and the samples are not excluded in the present analysis. The decision of whether to exclude these or not is beyond the scope of this first example, as numerical aspects are of primary concern. The concentration ranges were 6.88–34.40 mM for Co, 3.06–15.29 mM for Cr and 15.70–78.8 mM for Ni. The samples were measured in a range of 300–650 nm in 2 nm intervals on an HP 8452 DAD (Hewlett-Packard, Palo Alto, CA, USA) using a 1 cm quartz cuvette. For each sample, five replicate measurements were made by repeated randomized measurements of the samples. From these replicates the uncertainty of each measurement is calculated and used in a maximum likelihood PCA model assuming independent but different errors for each element. The loss function is illustrated graphically in Figure 2.

A PCA model is sought in which the scores are subsequently used for building a least squares regression model (hence a principal component regression (PCR) model). Several alternatives are tested here, mainly to illustrate the appropriateness of the algorithm by comparing to the earlier suggested ML-PCA algorithm given by Wentzell and co-workers [9,21], but also to show how simple it is to include centering in the maximum likelihood estimation with MILES, which is not the case for ML-PCA [9].

For illustration, the MILES algorithm for PCA with centering is given below. Note that this algorithm handles correlated errors, but as **W** is diagonal in this application, the weighted least algorithm of Kiers [1] could also be used.

Algorithm MILES-PCA

1. Initialize model $\mathbf{m}_0$, using centered LS-PCA model of the data, and set c:= 0.
2. $\mathbf{q} = \mathbf{m}_c + (1/\beta)\,\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{m}_c)$.
3. **T** and **P** are found as the PCA parameters when fitted to centered **Q**, i.e. to $\mathbf{Q} - \mathbf{1}\mathbf{n}^{\mathrm{T}}$, where **n** ($J \times 1$) holds the averages of the J columns of **Q** and where **Q** is the vector **q** arranged to the same size as the original data matrix.
4. $\mathbf{m}_{c+1} = \mathrm{vec}(\mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{1}\mathbf{n}^{\mathrm{T}})$.
5. c:= $c + 1$; go to step 2 until $\|\mathbf{m}_c - \mathbf{m}_{c-1}\|^2 / \|\mathbf{m}_{c-1}\|^2 \leq \varepsilon$.

In order to evaluate different PCR models, a leave-one-out cross-validation scheme was used where each sample was left out in turn. A PCA model was fitted and the scores were used in an ordinary multiple linear regression model for predicting Co(II), Cr(III) and Ni(II). Afterwards, the reference value of the left-out sample was predicted.

The following different PCA models were investigated, all with three principal components.

- A maximum likelihood PCA model was used where scores and centering parameters were estimated with MILES. Centering was included in the maximum likelihood fitting.
- A corresponding least squares PCA model was also used with ordinary centering.
- Finally, a maximum likelihood PCA model was estimated but with the data centered in an ordinary least squares sense. This model was fitted using both the MILES algorithm suggested here and the ML-PCA algorithm of Wentzell *et al.* [9].

The corresponding prediction results for the three components are shown in Table I as the root mean square error in cross-validation (RMSECV), which is defined as

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{I} e_{i_{\mathrm{CV}}}^2}{I}} \tag{23}$$

where $e_{i_{CV}}$ is the residual of the predicted valued of sample i predicted from a regression model built with the ith sample excluded. All models provide excellent results (correlations higher than 0.99), but the maximum likelihood results are never worse than the least squares results, suggesting that the use of maximum likelihood is feasible here. The results also show that MILES and the earlier algorithm suggested specifically for PCA give identical predictions as they should. The identical predictions do not prove that MILES provides the same PCA parameter estimates as ML-PCA. In order to verify that, a three-component PCA model was fitted 10 times to different random subsets of the 26 samples (15 samples in each subset). The parameters were estimated with both MILES and ML-PCA. The difference in fit was identical to minimally the sixth digit, also indicating that there are no significant problems with local minima solutions. The loadings obtained from MILES never differed more than maximally 0.0001% ($100\|\mathbf{P}^{\mathrm{MILES}} - \mathbf{P}^{\mathrm{ML-PCA}}\|/$

Table I. RMSECV values (M) for PCR model using three components. The first column shows the result for a maximum likelihood model including maximum likelihood centering. The second column is an ordinary least squares PCR. The third and fourth columns show the results using maximum likelihood estimation but ordinary least squares centering with the Wentzell and the MILES algorithm respectively. As expected, these two are identical

	MILES	LS	Wentzell (no ML offset)	MILES (no ML offset)
Cr	0.0897	0.0904	0.0894	0.0894
Ni	0.304	0.3464	0.3031	0.3031
Co	0.2931	0.3045	0.2934	0.2934

$\|\mathbf{P}^{\mathrm{ML-PCA}}\|$) from the loadings obtained from ML-PCA when correcting for possible rotational differences. Finally, it is noted that the inclusion of the centering within the maximum likelihood estimation is straightforward, although the significance of this with respect to predictions is limited here.

3.2. Signal processing—PARAFAC

An example is provided here on the usefulness of MILES in a different field, namely signal processing. The problem pertains to so-called 'blind' beamforming using receive antenna arrays [22]. The objective of blind beamforming is to reconstruct the emitted signal(s) and propagation parameters of radio waves propagating along multiple paths, each with its own attenuation and delay, without explicit knowledge of the propagation environment. In such problems the data arise out of band-limited radio signals, and the first step on the receiver side is to down-convert and lowpass filter the received data. This is done to filter off out-of-band interference. At the same time the filtered baseband signals are oversampled beyond the (minimum possible) Nyquist sampling rate in order to help resolve path delays and ensure that a PARAFAC model [5, 12, 22, 23] is appropriate. This oversampling of the filtered signal creates correlation of the noise samples along the oversampling mode, because the spectrum of the noise has been shaped by the frequency response of the lowpass filter.

The data are a $2 \times 4 \times 20$ array with additive white Gaussian noise. A two-component (two-ray) PARAFAC model is suitable for the data, which are generated according to

$$x_{ijk} = \sum_{f=1}^{2} a_{if} b_{jf} c_{kf} + e_{ijk}$$
$$i = 1,2, \quad j = 1,\ldots,4, \quad k = 1,\ldots,40 \qquad (24)$$

The first mode corresponds to receive antenna elements (two antennas are used), with the parameter a_{if} holding the gain for the ith antenna with respect to the fth signal. The second mode corresponds to symbol snapshots collected (four), given by the parameters b_{jf}, and the third mode to the number of samples taken per symbol interval, given by c_{kf}. The noise is held in e_{ijk} (Figure 3). The parameters are randomly drawn from a Gaussian (0,1) distribution and the residuals from a Gaussian (0,0.1) distribution. These noisy

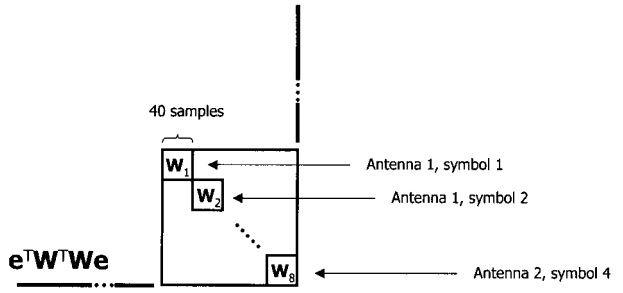

Figure 3. The loss function for the signal-processing data. The residual variation is independent from antenna to antenna and from symbol to symbol but is correlated within the third sample mode. Thus for each combination of receive antenna (first mode) and symbol snapshot (second mode) a characteristic error covariance matrix is obtained.

data are filtered by a five-sample moving average filter along the long oversampling mode to simulate the effect of the receive lowpass filter. This filtering step colors the noise spectrum according to a $\sin(x)/x$ pulse in the frequency domain, inducing noise correlation along the long mode.

For maximum likelihood estimation the error covariance matrix is estimated from 30 realizations of the noise. In Figure 4 the results from 100 runs are shown in terms of the signal-to-noise ratio (SNR) of the estimated symbols (sorted in size of maximum likelihood results). SNR is the common fidelity measure used for assessing the quality of the model in these applications and is defined by the true matrix $\mathbf{B}$ as well as the estimated $\hat{\mathbf{B}}$ as

$$\mathrm{SNR_B} = 20 \log_{10}\left(\frac{\|\mathbf{B}\|}{\|\mathbf{B} - \hat{\mathbf{B}}\|}\right) \qquad (25)$$

It is hence a measure of how well the loading matrix is recovered.

As can be seen, the maximum likelihood estimates are significantly better except for a few distinct cases. The generally higher SNR of the maximum likelihood estimates translates directly into better source/path localization and also reduced error rates in source signal recovery in the case of digital communication signals.

3.3. Fluorescence spectroscopy—handling scatter

3.3.1. Description of the problem

In fluorescence spectroscopy, scattering phenomena such as Rayleigh and Raman scatter are typically considered as noise, and often the areas where the scatter occurs are simply removed from the data beforehand or equivalently treated as missing data [23]. However, the interesting chemical information is sometimes situated in the areas where the scatter occurs and it is then not possible to treat these areas as missing data. It then becomes crucial to be able to handle and possibly separate the physical noise arising from scatter from the chemical information.

Baunsgaard [24] analyzed aqueous mixtures of four different fluorophores, namely L-phenylalanine, L-3,4-dihydroxyphenylalanine (DOPA), 1,4-dihydroxybenzene (hy-

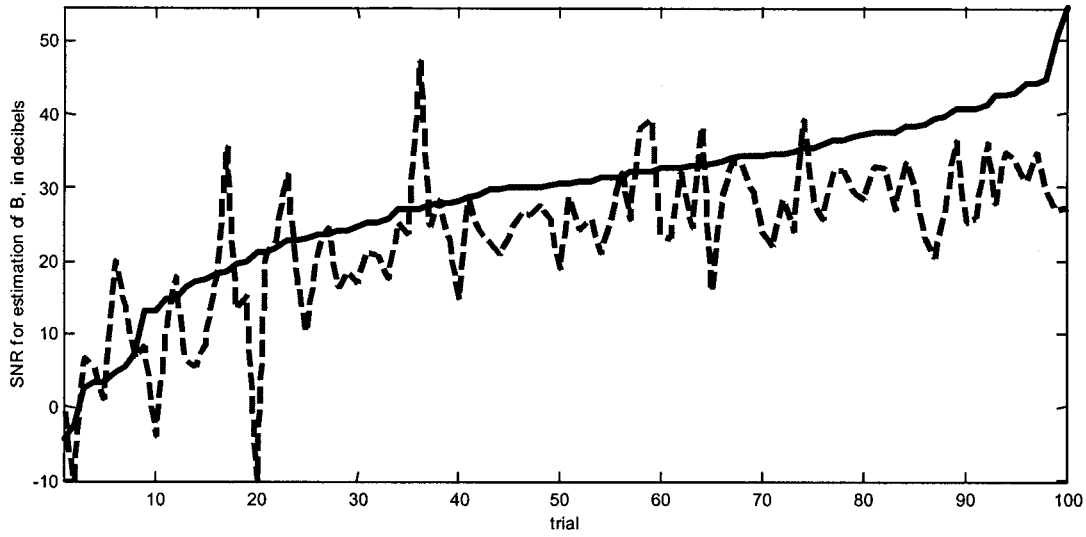

Figure 4. Signal-to-noise ratios (SNR) for 100 simulated models. The data are fitted by maximum likelihood PARAFAC (solid line) and ordinary least squares PARAFAC (broken line). The ML results in the figure are sorted in order of increasing SNR to allow easier visual comparison to the LS solution. As can be seen, the use of MILES provides better results in general.

droquinone) and L-tryptophan. Samples were prepared from stock solutions according to the design in Table II. Fluorescence excitation–emission landscapes of the 22 samples were obtained with a Perkin-Elmer LS50 B fluorescence spectrometer using excitation wavelengths between 200 and 350 nm and emission wavelengths between 200 and 750 nm. Excitation and emission monochromator slit widths were set to 5 nm and the scan speed was 1500 nm min^{-1}. In order to keep the data manageable and exclude irrelevant areas, subsets of the emission and excitation wavelengths were chosen. Thus the actual data used in the models contained excitations from 245 to 305 nm in 5 nm intervals and emissions from 246 to 436 nm in 5 nm intervals. A typical sample is shown in Figure 5 (left).

3.3.2. *A model of fluorescence data*

For dilute samples, fluorescence excitation–emission measurements ideally follow a trilinear PARAFAC model [25–28]. However, two types of problematic area exist in such data [25, 26, 29]. First of all, emission below excitation does not exhibit any fluorescence, and the intensity is simply zero. This part of the data does not necessarily follow the PARAFAC model, and therefore emission data below the excitation wavelength have to be set to missing or down-weighted such that the model does not incorrectly try to fit these zero values. One situation in which the trilinear PARAFAC model is not valid for emission below the excitation wavelength is when the excitation spectrum has non-zero values above the wavelength where the emission spectrum has non-zero values. In such a situation the outer product of the vectors holding the emission and excitation spectrum will be non-zero in areas of emission below excitation wavelengths. This contradicts the fact that physical measurements will be zero and thus illustrates that the PARAFAC model is not generally valid in this area. If not removed, this part of the data will therefore bias the estimated parameters towards zero. In Figure 5 a typical landscape is shown before and after removal of emission below excitation. The other problematic type of variation in such data is the so-called scatter ridges. Most significant in Figure 5 is the first-order Rayleigh scatter ridge that is seen in the rightmost part of the landscape [30]. It occurs approximately on the diagonal where the excitation equals the emission wavelength. This part of the data is problematic because it does not provide any chemical information, but only physical information that is not interesting with respect to the PARAFAC modeling. Further, such a ridge lying on a

Table II. Concentrations of four fluorophores in 22 samples (10^{-6}M)

Sample	Hydroquinone	Tryptophan	Phenylalanine	DOPA
1	17.00	2.00	4700	28.00
2	20.00	1.00	3200	8.00
3	10.00	4.00	3200	16.00
4	6.00	2.00	2800	28.00
5	0.00	0.00	5600	0.00
6	0.00	8.00	0	0.00
7	56.00	0.00	0	0.00
8	28.00	0.00	0	0.00
9	0.00	0.00	0	5.00
10	0.00	0.00	700	0.00
11	0.00	16.00	0	0.00
12	3.50	1.00	350	20.00
13	3.50	0.50	175	20.00
14	3.50	0.25	700	10.00
15	1.75	4.00	1400	5.00
16	0.88	2.00	700	2.50
17	28.00	8.00	700	40.00
18	28.00	8.00	350	20.00
19	14.00	8.00	175	20.00
20	0.88	8.00	1400	2.50
21	1.75	8.00	700	5.00
22	3.50	2.00	700	80.00

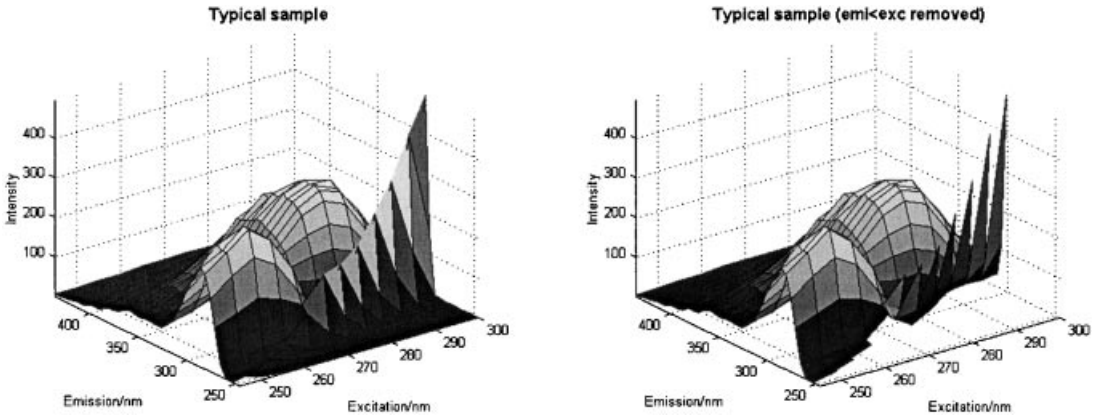

Figure 5. Left: a typical excitation–emission landscape. Note the diagonal Rayleigh scatter peak to the right. In the right plot, emission below excitation has been removed. Some scatter signal remains, but the bulk part of the data that does not follow the PARAFAC model has been removed.

diagonal is not possible to model efficiently e.g. by using an added PARAFAC component. The presence of scatter is cumbersome, especially when the chemical information appears close to the ridge. If this is not the case, it is sometimes possible to eliminate the scatter by setting the areas where the scatter occurs to missing or downweighting the elements [23, 31–34]. In addition to first-order Rayleigh scatter, other sources of scatter also occur. For these data a minor banded ridge of Raman scatter is observed for some of the samples. This type of scattering is less problematic in the sense that it usually is of a minor magnitude and because it can often be almost completely removed by subtracting measurements of the solvent without sample.

Two alternative PARAFAC models are tested for fitting these data. Both models are fitted with four components, one for each chemical analyte, with the aim of resolving the information for each analyte in each component. The alternative PARAFAC models are:

- least squares fitting to the raw data but eliminating emission below excitation;
- weighted least squares fitting using MILES, where emission below excitation is downweighted and weights are used to minimize the influence of scatter.

In order to define the weights in the MILES model, a simplified model of the scatter is used. This model is shown in Figure 6 (left). It consists of Gaussian curves generated as the sum of

$$r_{jk}^{\text{Rayleigh}} = h^{\text{Rayleigh}} \frac{e^{-[(\mu_j-\mu_k)/\sigma_{\text{Rayleigh}}]^2/2}}{\sigma_{\text{Rayleigh}}\sqrt{2\pi}} \tag{26}$$

and

$$r_{jk}^{\text{Raman}} = h^{\text{Raman}} \frac{e^{-\{[\mu_j-\mu_k(1.1+0.1j/J)]/\sigma_{\text{Raman}}\}^2/2}}{\sigma_{\text{Raman}}\sqrt{2\pi}} \tag{27}$$

The parameters of these distributions are μ_j (emission wavelength) and μ_k (excitation wavelength) and σ^{Rayleigh}, σ^{Raman}, h^{Rayleigh} and h^{Raman}, which are set to 8, 2, 25 and 0.5 respectively. These values were chosen rather arbitrarily to provide a reasonable visual resemblance to the observed sizes of the scatter peaks. In actual practice the width and height will change somewhat across the ridges, and the ridges will not be perfectly Gaussian shaped. However, it is anticipated that the approximation of the scatter is sufficiently good for the purpose of minimizing the negative effects of scatter.

The combined effect of scatter, measurement noise and downweighting of emission below excitation is shown in Figure 6 (right). This matrix is the sum of an i.i.d. term of magnitude standard deviation one, a missing data part of magnitude 1000 and the scatter model of Figure 6 (left). The absolute sizes of these three contributions are immaterial, whereas the relative sizes define the influence of the different residuals on the loss function. The weights used in MILES fitting are simply taken to be the inverse of the values in Figure 6 (right). Some comments on this fairly crude approach to defining the weights are appropriate.

- The same weights are applied to each sample. Although some differences in actual uncertainty may appear between samples, these are ignored here.
- The model fitted with the above-defined weights is not a maximum likelihood estimate. First of all, the scatter has a bias part which must be eliminated in order to be able to fit the model in a maximum likelihood sense, and secondly, the error estimates are not based on actual estimates of uncertainty.
- Eliminating bias by centering, though, is not feasible in this case. In practice the bias can be removed by centering the data across samples [2], but this will lead to overfitting, because a huge number of averages need to be calculated which are not strictly necessary.
- Improving the error estimates based e.g. on actual replicate measurements is avoided here in favor of simpler estimates. Estimating actual error variances and covariances from empirical data can lead to quite noisy estimates, which will possibly even have less resemblance to the true error covariances than simply using the implicit i.i.d. assumptions of a least squares approach.
- No use is made of off-diagonals in the weighting matrix (**W** is diagonal as in the weighted least squares loss function in

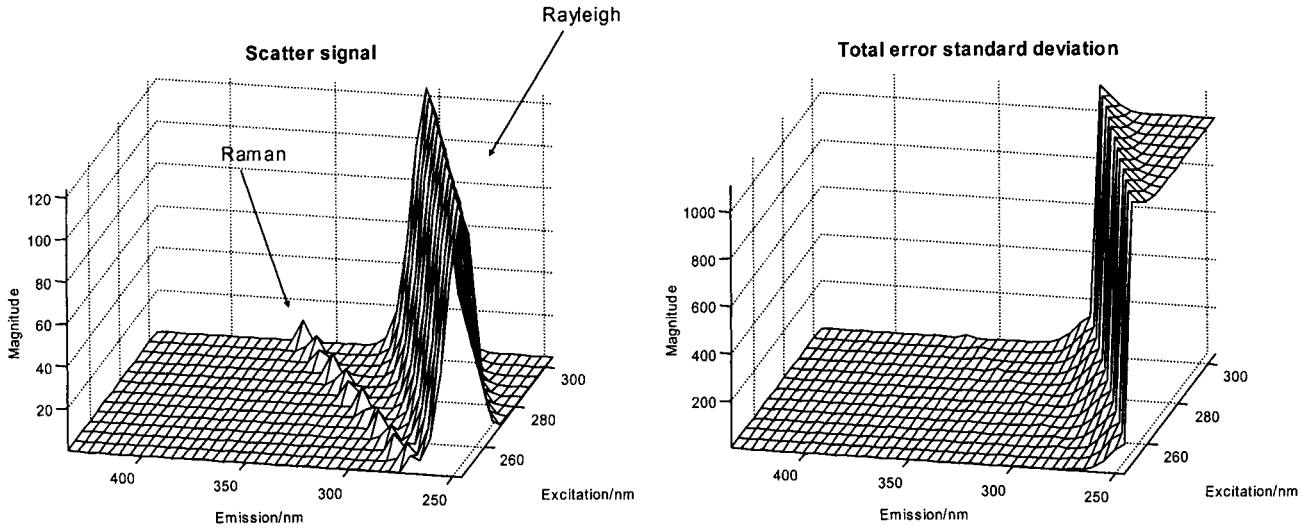

Figure 6. Left: an idealized representation of first-order Rayleigh and Raman scatter. The large Rayleigh trace appears around the diagonal where the excitation wavelength equals the emission wavelength, while the Raman scatter has a lower intensity and moves along a different diagonal. To the right, the total error 'standard deviation' for one sample is given. It is the sum of the scatter error from the left plot, the i.i.d. error (standard deviation one for every element) and a very high value (1000) for the non-chemical part of the data.

Figure 2). This is justified by the adequacy of the results obtained without off-diagonals and the computational simplicity of the associated matrix inversions when off-diagonals are left out.

3.3.3. *Qualitative results*

In Figure 7 the resulting estimated loading parameters are shown for two competing four-component PARAFAC models. It is evident that there is a discrepancy in one of the estimated emission spectra and also in one of the excitation spectra.

The shape of the 260 nm artifact peak for the least squares emission component 3 (solid line) is naturally related to the Rayleigh scatter, but, equally importantly, the relatively high magnitude of the peak occurs because of the pattern of missing data. Looking at the model approximation from only component 3, i.e. the outer product of the emission and excitation loading, it is seen that although the 260 nm peak appears large in the emission loading, its contribution to the

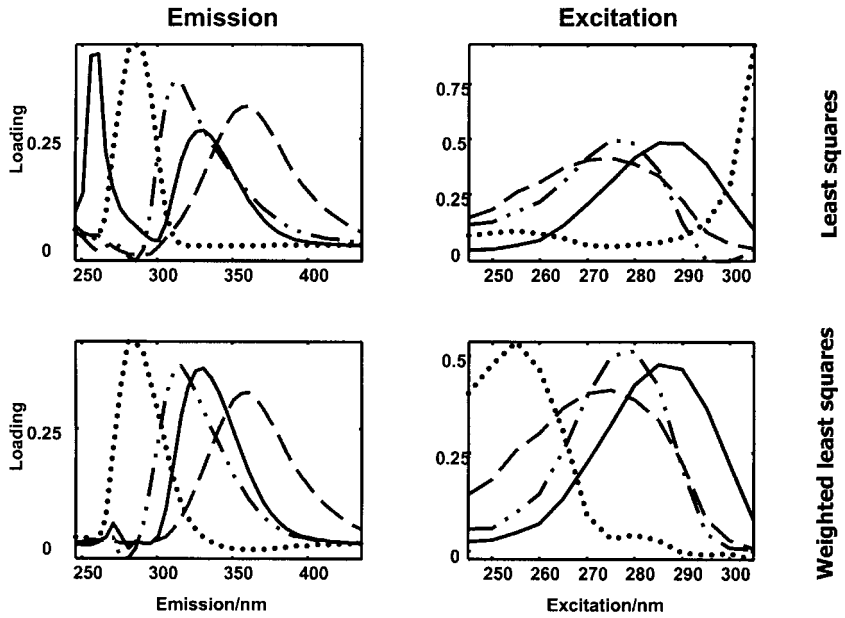

Figure 7. Four components estimated with least squares (top) and weighted least squares (bottom) approaches. Component 3 (solid line) has a peculiar peak in the low-emission part in the least squares approach. Component 2 (dotted line) has an incorrect excitation maximum above the corresponding emission maximum in the least squares model.

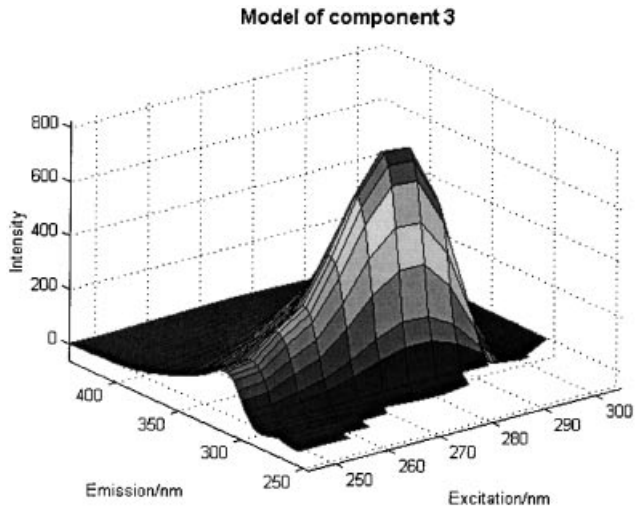

Figure 8. The landscape of component 3 (outer product of emission and excitation loading) for the least squares fitted model. The emission wavelengths below excitation have been set to missing and it is seen that the apparent high scatter peak in the emission loading is almost absent from this landscape because of the missing data.

model is quite moderate (Figure 8) when the area without fluorescence information is taken into account. Thus the apparent large size, and hence importance, of the scatter peak is an artifact of the special data structure. In actual practice the magnitude of the 260 nm peak can be changed with little associated change in fit. Such artifacts are often observed when modeling fluorescence data, even if emission slightly above the excitation wavelength has also been removed to minimize the effect of scatter [29].

The high value of the least squares excitation component 2 (dotted line) is also related to the Rayleigh scatter. Owing to the low signal from this component and a relatively high amount of Rayleigh scatter, the component is primarily reflecting the Rayleigh trace rather than the chemical variation. When fitting the model with the MILES procedure, the artifact is not seen and the overall estimated spectra seem more reasonable. Hence the use of the WLS fitting seems warranted in this case.

As an aside, fitting a least squares model to data where non-fluorescent zeros from emission below excitation and scatter are retained (i.e. not treated as missing), the results are sometimes similar to the ones obtained here with MILES fitting. This *indicates* that the two approaches are equally good. However, the reason why the MILES approach works is because it handles the special characteristics and deviations from the ideal 'i.i.d. least squares' conditions in a direct and reasonable way. On the other hand, when fitting the data with a large number of 'incorrect' zeros, these zeros act in the model in a similar way to ridge parameters in a regression model. Because the model is forced to describe the zeros, the parameters are forced towards zero. Further, because the scatter peaks in the estimated spectra only describe minor variation in the data, the gain in fit by describing the incorrect zeros supersedes the loss in fit by not describing the scatter peak. This is an *ad hoc* approach and can be expected to bias the solution in many cases. Indeed, for these data, this approach is not feasible. This is further corroborated by means of the quantitative results that follow.

3.3.4. *Quantitative results*

In order to substantiate the adequacy of the MILES approach, the following resampling approach was adopted.

1. Ten of the 22 samples are randomly chosen.
2. A four-component PARAFAC model is fitted to these 10 samples.
3. The 10 concentrations of the four analytes are predicted by the PARAFAC scores using multiple linear regression (no offset).
4. For each analyte, R^2 is computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I} y_i^2}$$

   where y_i is the concentration in the ith sample and $\hat{y}_i$ is the corresponding prediction.
5. This procedure is repeated 100 times, yielding 400 R^2 values.

The R^2 statistic provides a measure of the fraction of variance explained [35] and is expressed as a percentage in the results that follow. The closer this percentage is to 100%, the better the model fits the data. Note that the predictions of $\mathbf{y}$ are based on a no-intercept model.

The following alternative loss functions and models were evaluated.

A. Least squares fit to raw data (no missing data, hence emission below excitation and scatter are retained (see Figure 5 left)).
B. Least squares fit to data (setting emission below excitation to missing (see Figure 5 right)).
C. Weighted least squares fitting (as specified above).

Least squares fitting without handling the area where fluorescence does not occur is performed in A in order to test if this is a feasible approach as discussed above. In B, ordinary least squares fitting is used, but setting emission below excitation to missing. The missing data are handled by expectation maximization [25]. This idea of eliminating or downweighting the problematic areas is the approach most often used for multi-way modeling of fluorescence data. In this case it would likely be possible to push the limit for setting elements to missing a bit. Setting emission slightly above the excitation wavelength to missing can help remove even more of the problematic scatter. However, as this will also eliminate chemical information, which is potentially important for example for low-wavelength-emitting analytes, this is not pursued here.

In Figure 9 the results obtained from 100 runs for each of the three alternative models are shown. Handling the missing data (B) is slightly better than disregarding the non-fluorescent parts (A), and using the MILES approach provides the best results by far (C). This result, of course, is simply a manifestation of the results outlined in the

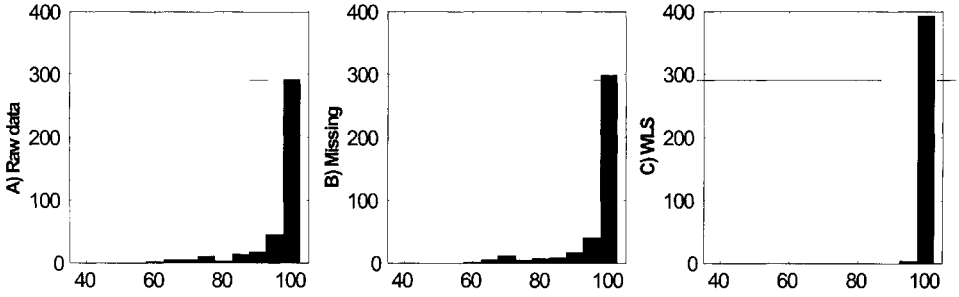

Figure 9. Histograms of 400 values of $R^2 \times 100$ obtained from the three different models.

discussion of the qualitative results. The results can be further understood e.g. by looking at the actual emission loadings estimated by the three models in all 100 cases. From Figure 10 it is easily seen that only the MILES models (C) are able to accurately determine the spectra in all cases. For the two alternative models the correct estimates are only obtained occasionally, whereas in many cases incorrect models are obtained, reflecting the scattering phenomena.

The model of the scatter phenomena used for defining the weights has been chosen in an *ad hoc* fashion based on visual assessment of the model. In order to verify that the results obtained with MILES are robust towards slightly different weights, two alternative models were investigated, one where the scatter ridges are incorrectly moved as much as 10 nm away in the emission direction from the correct diagonal where emission equals excitation, and one where the Rayleigh trace is only half the width of the former. As can be seen in Figure 11, the resulting PARAFAC loadings are almost identical. Although minor deviations appear in the estimated spectra, these are insignificant compared to the differences observed in the least squares model (Figure 7). The results do indicate that some optimization of the scatter model may be achieved, but, more importantly, it is verified that the model is robust against minor changes in the error model. Hence any reasonable model of the scatter will help in correcting the problems encountered in the least squares approach.

4. CONCLUSION

A new and general algorithm has been developed for maximum likelihood or weighted least squares estimation. It is applicable to any situation where a least squares algorithm exists, and can be implemented as a simple iterative procedure. This algorithm makes it simple e.g. to make maximum likelihood fitting algorithms in situations where it would otherwise be difficult to come up with a suitable algorithm. One such example is to fit a bilinear PCA model including centering. The centering part is not easily handled with other maximum likelihood PCA algorithms,

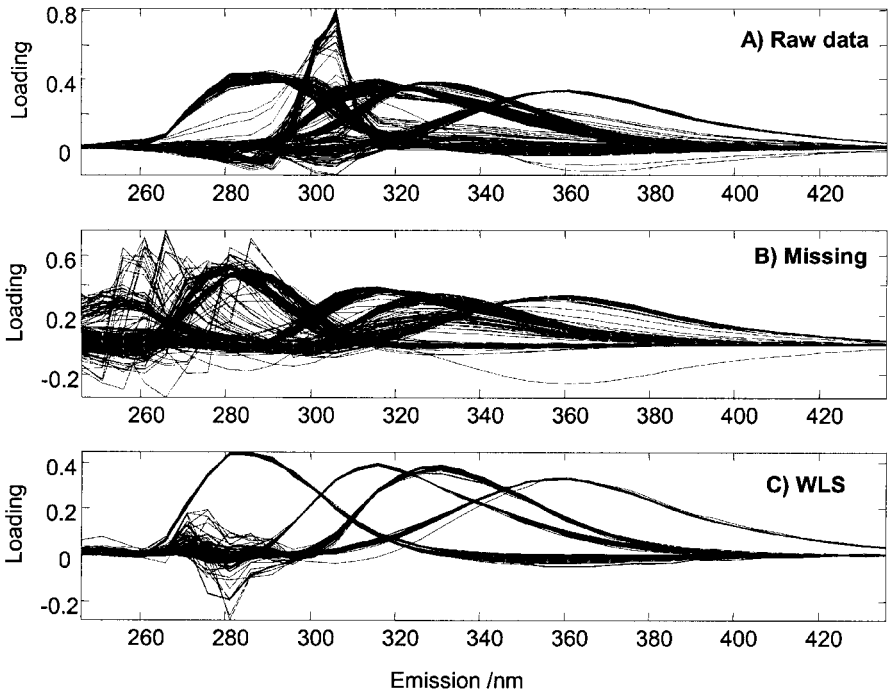

Figure 10. Emission loadings obtained from a least squares model of raw data (top), a least squares model handling missing data (middle) and a weighted least squares model (bottom). For each type of model the results from the 100 refitted models are superimposed.

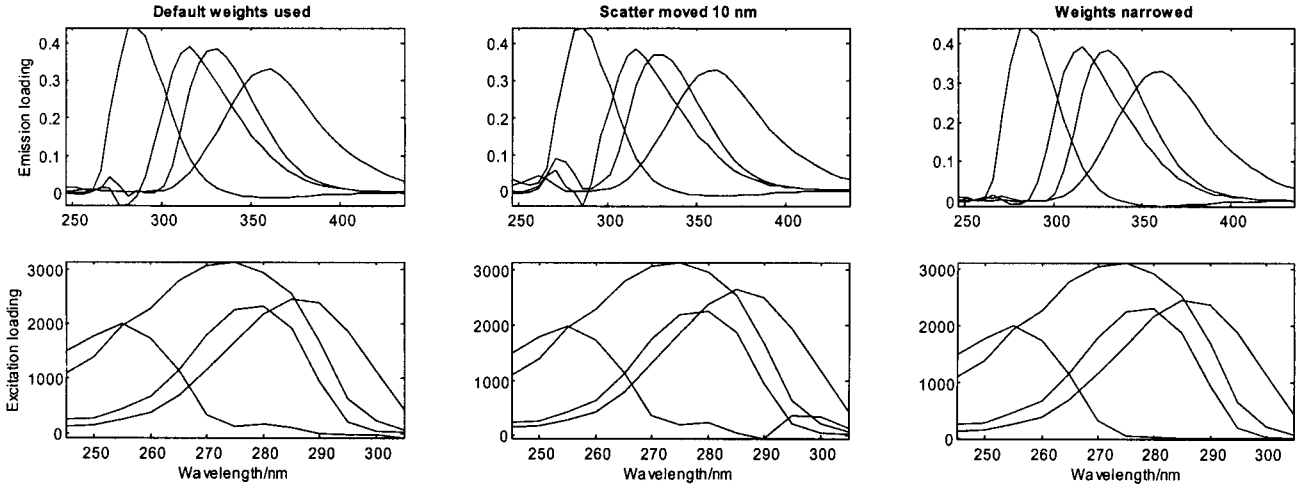

Figure 11. Results from PARAFAC using three alternative scatter models. Leftmost, the model used so far; then a model where the scatter ridges are (incorrectly) moved 10 nm away from the correct diagonal in the emission direction; and finally a model where the Rayleigh trace is half the width. The resulting emission and excitation loadings are shown.

but for MILES the structure of the model is immaterial as long as a least squares algorithm exists. The algorithm has been applied to different data sets to show its applicability. These examples include PCA and PARAFAC. It has been shown that very accurate knowledge of the error covariance structure is not mandatory for the beneficial use of the weighted least squares principle. Sometimes, though, maximum likelihood estimation is disregarded in favor of, for example, least squares approaches. Indeed, if the residuals have quite similar variances and are imprecisely determined from few replicates, it is conceivable that the use of such uncertain variances can lead to poor results. The difference between maximum likelihood and least squares fitting, though, becomes less pronounced if the errors are small compared to the signal (which is often the case in e.g. near-infrared spectroscopy).

The main results of the presented work are thus as follows.

- MILES is a general algorithm that can be used to replace a number of specialized algorithms such as the algorithms developed by Wentzell, Kiers, etc.
- The spectroscopic example basically shows that the MILES algorithm provides the same results as current ML-PCA algorithms and that offsets are easily included in the maximum likelihood fitting.
- The signal-processing example shows that MILES handles residual covariance e.g. for fitting PARAFAC models. No other PARAFAC algorithm currently does this.
- The fluorescence example provides an example of how the weighted least squares approach of MILES can be used for solving a well-known problem in modeling fluorescence data.

Acknowledgements

R. Bro gratefully acknowledges support provided by the LMC (Center for Advanced Food Studies) and Frame program AQM (Advanced Quality Monitoring in the Food Production Chain), as well as the EU (European union) under project GRD1-1999-10337, NWAYQUAL. N. D. Sidiropoulos gratefully acknowledges support provided by NSF (National Science Foundation) grants 0096164 and 0096165. The data for and m-file performing ML-PCA were obtained from http://www.dal.ca/~pdwentze/index.html.

APPENDIX

For the second term in Equation (20) it holds that

$$\beta\ (\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}(\mathbf{m}-\mathbf{m}_c)=\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}+\beta\ \mathbf{m}_c^{\mathrm{T}}\mathbf{m}_c-2\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}_c \tag{28}$$

where $\mathbf{m}_c^{\mathrm{T}}\mathbf{m}_c$ is a constant. The third term in Equation (20) can be rearranged as

$$\begin{aligned}2\ (\mathbf{m}-\mathbf{m}_c)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}_c)=2\ (&\mathbf{m}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}-\mathbf{m}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c\\&-\mathbf{m}_c^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}+\mathbf{m}_c^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c)\end{aligned} \tag{29}$$

Therefore Equation (20) can be written as

$$\begin{aligned}\sigma_{\mathrm{maj}}(\mathbf{m}|\mathbf{x},\mathbf{W},\mathbf{m}_c)=\alpha+\beta\ &\mathbf{m}^{\mathrm{T}}\mathbf{m}+\beta\ \mathbf{m}_c^{\mathrm{T}}\mathbf{m}_c-2\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}_c\\&-2\ (\mathbf{m}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}-\mathbf{m}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c\\&-\mathbf{m}_c^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}+\mathbf{m}_c^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c)\end{aligned} \tag{30}$$

Defining the constant term $\delta=\alpha+\beta\ \mathbf{m}_c^{\mathrm{T}}\mathbf{m}_c+2\mathbf{m}_c^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}-2\mathbf{m}_c^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c$ leads to

$$\begin{aligned}\sigma_{\mathrm{maj}}(\mathbf{m}|\mathbf{x},\mathbf{W},\mathbf{m}_c)&=\delta+\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}-2\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}_c-2\ \mathbf{m}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}\\&\quad+2\ \mathbf{m}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c\\&=\delta+\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}-2\ \mathbf{m}^{\mathrm{T}}(\beta\ \mathbf{m}_c+\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}\\&\quad-\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{m}_c)\\&=\delta+\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}-2\ \mathbf{m}^{\mathrm{T}}[\beta\ \mathbf{m}_c+\mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x}-\mathbf{m}_c)]\\&=\delta+\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{m}-2\beta\ \mathbf{m}^{\mathrm{T}}\mathbf{q}\end{aligned} \tag{31}$$

REFERENCES

1. Kiers HAL. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika* 1997; **62**: 251–266.
2. Bro R and Smilde AK. Centering and scaling in component analysis. *J. Chemometrics* in press.
3. Gabriel KR. Least squares approximation of matrices by additive and multiplicative models. *J. R. Statist. Soc. B* 1978; **40**: 186–196.
4. Kruskal JB. Some least squares theorems for matrices and N-way arrays. *Manuscript*, Bell Laboratories, Murray Hill, NJ, 1977.
5. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multimodal factor analysis. *UCLA Working Papers Phonet* 1970; **16**: 1–84.
6. Martens H, Høy M, Wise BM, Bro R and Brockhoff PMB. GLS preprocessing of multivariate data. *J. Chemometrics* in press.
7. Paatero P and Tapper U. Analysis of different modes of factor analysis as least squares problems. *Chemometrics Intell. Lab. Syst.* 1993; **18**: 183–194.
8. Paatero P and Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 1994; **5**: 111–126.
9. Wentzell PD, Andrews DT, Hamilton DC, Faber NM and Kowalski BR. Maximum likelihood principal component analysis. *J. Chemometrics* 1997; **11**: 339–366.
10. Judge GG, Griffifths WE, Carter Hill R, Lütkepohl H and Lee TC. *The Theory and Practice of Econometrics*. Wiley: New York, 1985.
11. Kay S. *Statistical Signal Processing, Part I: Estimation Theory*. Prentice-Hall: Englewood Cliffs, NJ, 1993.
12. Golub GH and van Loan CF. *Matrix Computations*. Johns Hopkins University Press: Baltimore, MD, 1989.
13. De Lathauwer L, de Moor B and Vandewalle J. An introduction to independent component analysis. *J. Chemometrics* 2000; **14**: 123–149.
14. Kroonenberg PM. *Three-mode Principal Component Analysis. Theory and Applications*. DSWO Press: Leiden, 1983.
15. Groenen PJF, Mathar R and Heiser WJ. The majorization approach to multidimensional scaling for Minkowski distances. *J. Classif.* 1995; **12**: 3–19.
16. Heiser WJ. Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. In *Recent Advances in Descriptive Multivariate Analysis*, Krzanowski WJ (ed.). Oxford University Press. 1995; 157–189.
17. Kiers HAL. Majorization as a tool for optimizing a class of matrix functions. *Psychometrika* 1990; **55**: 417–428.
18. Kiers HAL and ten Berge JMF. Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika* 1992; **57**: 371–382.
19. Heiser WJ. Correspondence analysis with least absolute residuals. *Comput. Statist. Data Anal.* 1987; **5**: 337–356.
20. Liu XQ and Sidiropoulos ND. Cramer–Rao lower bounds for low-rank decomposition of multidimensional arrays. *IEEE Trans. Signal Processing* 2001; **49**: 2074–2086.
21. Wentzell PD and Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemometrics Intell. Lab. Syst.* 1999; **45**: 65–85.
22. Sidiropoulos ND and Liu XQ. Identifiability results for blind beamforming in incoherent multipath with small delay spread. *IEEE Trans. Signal Processing* 2001; **49**: 228–236.
23. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
24. Baunsgaard D. Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes. *Royal Veterinary and Agricultural University*, Frederiksberg, 1999.
25. 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).
26. Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics Intell. Lab. Syst.* 1999; **46**: 133–147.
27. Leurgans SE and Ross RT. Multilinear models: applications in spectroscopy. *Statist. Sci.* 1992; **7**: 289–319.
28. Ross RT and Leurgans SE. Component resolution using multilinear models. *Methods Enzymol.* 1995; **246**: 679–700.
29. Bro R and Sidiropoulos ND. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics* 1998; **12**: 223–247.
30. Ewing GW. *Instrumental Methods of Chemical Analysis*. McGraw-Hill: New York, 1985.
31. Bro R and Heimdal H. Enzymatic browning of vegetables. Calibration and analysis of variance by multiway methods. *Chemometrics Intell. Lab. Syst.* 1996; **34**: 85–102.
32. Heimdal H, Bro R, Larsen LM and Poll L. Prediction of polyphenol oxidase activity in model solutions containing various combinations of chlorogenic acid, (−)-epicatechin, O_2,CO_2, temperature and pH by multiway analysis. *J. Agric. Food Chem.* 1997; **45**: 2399–2406.
33. Jiji RD and Booksh KS. Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation–emission matrix fluorescence spectra. *Anal. Chem.* 2000; **72**: 718–725.
34. Wentzell PD, Nair SS and Guy RD. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Anal. Chem.* 2001; **73**: 1408–1415.
35. Draper NR and Smith H. *Applied Regression Analysis*. Wiley: New York, 1981.