

MAXIMUM LIKELIHOOD PRINCIPAL COMPONENT ANALYSIS

PETER D. WENTZELL,¹ DARREN T. ANDREWS,¹ DAVID C. HAMILTON,² KLAAS FABER³ AND
BRUCE R. KOWALSKI³

¹ Trace Analysis Research Centre, Department of Chemistry, Dalhousie University, Halifax, Nova Scotia B3H 4J3,
Canada

² Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 3J5,
Canada

³ Center for Process Analytical Chemistry, University of Washington, Seattle, WA 98195, U.S.A.

SUMMARY

The theoretical principles and practical implementation of a new method for multivariate data analysis, maximum likelihood principal component analysis (MLPCA), are described. MLPCA is an analog to principal component analysis (PCA) that incorporates information about measurement errors to develop PCA models that are optimal in a maximum likelihood sense. The theoretical foundations of MLPCA are initially established using a regression model and extended to the framework of PCA and singular value decomposition (SVD). An efficient and reliable algorithm based on an alternating regression method is described. Generalization of the algorithm allows its adaptation to cases of correlated errors provided that the error covariance matrix is known. Models with intercept terms can also be accommodated. Simulated data and near-infrared spectra, with a variety of error structures, are used to evaluate the performance of the new algorithm. Convergence times depend on the error structure but are typically around a few minutes. In all cases, models determined by MLPCA are found to be superior to those obtained by PCA when non-uniform error distributions are present, although the level of improvement depends on the error structure of the particular data set. © 1997 by John Wiley & Sons, Ltd.

Journal of Chemometrics, Vol. 11, 339–366 (1997) (No. of Figures: 4 No. of Tables: 5 No. of References: 22)

KEY WORDS principal component analysis; maximum likelihood; measurement errors; multivariate analysis; near-infrared spectroscopy; errors in variables

1. INTRODUCTION

In general the primary goal of chemometrics is to develop and utilize models for chemical measurements. Principal component analysis (PCA) has been perhaps the most powerful tool of the chemometrician in this regard. Initially employed by statisticians to describe the variance and covariance of random variables, PCA is more commonly used in chemometrics to describe deterministic relationships among variables, especially in cases where a high degree of collinearity exists. In this context the advantage of PCA is that it allows multivariate data to be represented by a smaller number of variables, called principal components, factors or latent variables. In applications such as mixture analysis and curve resolution the object is to develop a p -dimensional linear model (i.e. a p -dimensional hyperplane) to describe the data within experimental error. In this case p is sometimes called the chemical rank, pseudorank or true rank of the data set to distinguish it from the mathematical rank, which is nearly always maximized owing to the presence of experimental error. The chemical rank is typically related to the number of underlying chemical factors or chemical components present and in this paper will be referred to simply as the 'rank'. Although there are other applications of PCA, such as dimensionality reduction and preprocessing, this paper will consider the

Correspondence to: Peter D. Wentzell.

CCC 0886-9383/97/040339-28 \$17.50

© 1997 by John Wiley & Sons, Ltd.

Received 2 July 1996

Accepted 24 October 1996

technique from a modeling perspective.

PCA has been very successfully applied to modeling in chemistry, as evidenced by the large number of papers on mixture analysis and related topics. The objective of the chemist is generally twofold: (i) to determine the correct form of the model (the rank p as well as the presence of any model offsets); (ii) to obtain the best estimate of parameters associated with the model (in the form of eigenvectors, scores, etc.). Usually the hyperplane described by the model is of more interest than the eigenvectors themselves, since it typically describes the space containing the real factors, such as pure component spectra. Unfortunately, PCA is often not an optimal procedure for the estimation of model parameters and can lead to poor models in certain cases. Of course, there are a variety of optimization criteria to be considered when evaluating parameter estimation methods (e.g. robustness, bias and variance of the estimators) and none is universally the best. One widely used approach is to employ a maximum likelihood criterion. Simply put, for a given p -dimensional model the maximum likelihood solution for the model parameters is the one that is most likely to give rise to the observed measurements. Maximum likelihood estimates are generally recognized as having desirable statistical properties and their use has become commonplace.¹ For example, ordinary least squares is a maximum likelihood method when measurement errors in the x -variable(s) are negligible and the errors in y are independent and normally distributed. Likewise, PCA can be considered to be a maximum likelihood method if all measurement error standard deviations have the same normal distribution (i.e. independent and identically distributed (i.i.d.)). This is one reason for the widespread use of PCA for modeling applications. When minor variations from the assumptions for maximum likelihood estimation are observed, PCA can still be useful, but when the violations become large, it becomes ineffective. This has been remedied in part by incorporating various scaling techniques to reduce the data to i.i.d., but this will not work in the general case.² A maximum likelihood method which is more general in its approach to modeling is needed.

This paper establishes the theoretical foundations for maximum likelihood principal component analysis (MLPCA). MLPCA is an errors-in-variables modeling method in that it accounts for measurement errors in the estimation of model parameters. It is an optimal modeling method in a maximum likelihood sense for functional models with no errors in the model equations. The method is first presented in terms of a classical measurement error regression model and then transformed to principal component space to provide a closer relationship with PCA and a more tractable formulation. The mathematical aspects of the algorithm are described in detail to allow the principles to be readily applied.

A number of assumptions have been made in the development of the MLPCA method and these should be made clear from the outset. First, it is assumed that there is a true underlying p -dimensional model for the data (possibly with the inclusion of row and column offsets). Second, deviations of the measurements from this model are the result only of random measurement errors (no model errors or outliers). Third, these random errors are normally distributed around the true measurements with known standard deviations and covariance structure. Although the assumption of normally distributed errors may be violated in practice and does not strictly preclude the application of the maximum likelihood principle, mathematical tractability demanded that this assumption be made for the development of this algorithm and limits its generality to a small degree. However, correlated errors can be accommodated by the method as long as the error covariance matrix can be estimated. Finally, as noted, the method assumes that the true values of the measurement standard deviations (or error covariance matrices) are known, while in practice only estimates of these values are normally available. Although this will no doubt affect the quality of the parameter estimates, it does not affect the development of the principles here or the utility of the approach. It is the premise of this work that a modeling method which includes some knowledge of the measurement errors, however incomplete, is better than one which includes no information other than its implicit assumptions.

2. BACKGROUND

The need for some form of PCA which is weighted according to measurement errors has been recognized for some time. Informally this has been addressed through a variety of scaling procedures, such as range scaling and autoscaling as well as more elaborate schemes.²⁻⁵ Under the right conditions these can reduce PCA to a maximum likelihood method. However, Paatero and Tapper² showed that for this to happen, the matrix of measurement standard deviations needs to be of rank one. If this condition is not met, as it will not be in the general case, any weighting scheme developed will be suboptimal.

A number of researchers have developed methods to more rigorously incorporate measurement errors into the modeling process.^{6,7} This is normally done by minimizing the usual weighted residual sum of squares in accordance with some p -dimensional model. Mathematically, if \mathbf{X} is an $m \times n$ data matrix, this corresponds to minimization of

$$S^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ij} - \hat{x}_{ij})^2}{\sigma_{ij}^2} \quad (1)$$

where \hat{x}_{ij} corresponds to the estimated value of the measurement. In the general case where there are no offsets in the model, this is given by

$$\hat{\mathbf{X}} = \mathbf{AB} \quad (2)$$

where \mathbf{A} is $m \times p$ and \mathbf{B} is $p \times n$. By analogy to PCA, \mathbf{A} and \mathbf{B} correspond to score and loading matrices, but in equation (2) the individual vectors are not required to be orthogonal. A variety of methods have been devised to obtain \mathbf{A} and \mathbf{B} through minimization of equation (1) and these differ largely in their representation of the problem, the constraints applied to the solution and their approach to the non-linear optimization. Gabriel and Zamir⁶ describe a method based on 'criss-cross regressions' as a means to obtain lower-rank approximations of the matrix \mathbf{X} . Paatero and Tapper⁷ have described what they call 'positive matrix factorization' (PMF) and have applied this to environmental problems. In addition to satisfying the minimization criterion, PMF also requires \mathbf{A} and \mathbf{B} to be positive.

In this work, MLPCA is presented as an alternative to the above methods. Although MLPCA is similar to these methods in that it seeks to minimize equation (1), it has some important differences. First, it is formulated in terms of singular value decomposition (SVD), which is a very common method for implementing PCA. Second, the standard MLPCA algorithm consists of an alternating least squares procedure which is robust, easy to implement and very efficient compared with conventional gradient search methods. Finally, unlike the methods described in the preceding paragraph, which require measurement errors to be independent, MLPCA allows the inclusion of error covariance in virtually any form.

The MLPCA method described here should not be confused with maximum likelihood common factor analysis (MLCFA) that appears frequently in the literature outside of chemistry. Although the terms are often used interchangeably by chemists, PCA and factor analysis are distinctly different approaches to multivariate analysis.⁸ The principles of MLCFA were originally developed by Lawley and Maxwell⁸ and later employed in programs such as LISREL.^{9,10} More recently, MLCFA has appeared in the chemical literature with claims that it performs better than PCA.^{11,12} However, MLCFA was developed with the intention of finding structural models for random variables. As such, it estimates covariance matrices for random variables and does not generally use information about measurement errors.

Another errors-in-variables method that has become popular recently is total least squares (TLS).¹³

This method uses SVD for the purpose of developing a regression model and is similar to MLPCA in some ways. However, it is less general in its ability to obtain maximum likelihood estimates of model parameters. To our knowledge, the method described in this work is unique in its approach to PCA model estimation.

The objective of the work presented here is to develop the MLPCA approach in a manner consistent with the PCA formulation and present algorithms which are computationally practical. A complete analysis of the statistical properties of the method is beyond the scope of this treatment, but examples are presented to validate the method and demonstrate some of its features. Additional applications will be presented in future work.

3. THEORY

The development of the theoretical aspects of MLPCA is presented here in four subsections. First, the parametric models are developed and extended to a PCA framework and a strategy for gradient optimization of the model parameters is discussed. In the second subsection a more efficient optimization procedure based on an alternating least squares approach is described. This procedure assumes that the model contains no intercept terms and the measurements have uncorrelated errors. This algorithm will be referred to as the 'standard' MLPCA algorithm since it represents the simplest case. The more general case which accommodates correlated errors is discussed in Section 3.3. Finally, as an analog to mean centering in traditional PCA, the incorporation of intercept terms into the MLPCA procedure is treated in Section 3.4.

3.1. MLPCA with no intercept terms

Starting with the $m \times n$ matrix of measurements, \mathbf{X} , the MLPCA problem can be regarded simply as one of finding the equation for the optimum p -dimensional hyperplane to fit n points in the m -dimensional row space or, alternatively, m points in the n -dimensional column space. In the analysis presented here, the former approach is used, but it will become apparent that this is not important. Maximum likelihood model estimation is an iterative two-step procedure. First, for a set of given hyperplanar model parameters (i.e. slopes) the maximum likelihood estimates for the points (the column vectors of the observed data matrix \mathbf{X}) are found. These are then used to calculate the objective function in equation (1) (or an analogous equation). In the second step the model parameters are adjusted in an attempt to minimize S^2 . The new model parameters are used to calculate new maximum likelihood estimates of the points and a new S^2 and the process continues until the objective function is minimized. Thus there are two problems to address: (i) how to calculate the maximum likelihood estimates for a given set of model parameters; (ii) how to optimize the model parameters. These problems are treated in order.

In accordance with the assumptions stated earlier, each column of the data matrix \mathbf{X} can be considered to represent a point in the m -dimensional row space, with the true measurements corrupted by normally distributed errors:

$$\mathbf{x} = \mathbf{x}^0 + \boldsymbol{\varepsilon} \quad (3)$$

Here \mathbf{x} is a column vector of \mathbf{X} , \mathbf{x}^0 represents the error-free column vector and $\boldsymbol{\varepsilon}$ is the vector of measurement errors, which has an error covariance matrix

$$\boldsymbol{\Psi} = \text{cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) \quad (4)$$

where ' $E(\cdot)$ ' denotes an expectation value. Note that each column of \mathbf{X} can have a different error covariance matrix. In the development of an MLPCA model it is assumed that the error-free

measurements lie on a p -dimensional hyperplane that can be modeled by a set of parametric equations with p independent variables. The independent variables for these parametric equations will be arbitrarily chosen to be the first p rows of \mathbf{X} . For example, the parametric equations for a two-dimensional (planar) model in a four-dimensional space would be

$$x_3^o = a_{31}x_1^o + a_{32}x_2^o, \quad x_4^o = a_{41}x_1^o + a_{42}x_2^o \tag{5}$$

where the a s are the model parameters and the x^o s are the error-free measurements. In matrix form the equation for the observation is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \begin{bmatrix} x_1^o \\ x_2^o \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} \tag{6}$$

or in general

$$\mathbf{x} = \mathbf{A}\mathbf{x}_p^o + \boldsymbol{\varepsilon} \tag{7}$$

where \mathbf{x}_p^o is the vector containing the first p elements of \mathbf{x}^o . In this equation, \mathbf{A} is the $m \times p$ matrix of model coefficients (slope parameters) and the upper $p \times p$ submatrix of \mathbf{A} is the identity matrix. Our problem here is to find the best estimate for \mathbf{x}_p^o given the vector of observations, \mathbf{x} , a matrix of estimated model coefficients, $\hat{\mathbf{A}}$, and an error covariance matrix $\boldsymbol{\Psi}$. Note that this problem is similar in form (but different in objective) to the classical regression problem, which uses the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} \tag{8}$$

In this equation, \mathbf{y} and \mathbf{X} are the observations, $\boldsymbol{\beta}$ is the regression vector and $\boldsymbol{\delta}$ is the measurement error vector for \mathbf{y} . In the regression case we assume no errors in \mathbf{X} and try to estimate $\boldsymbol{\beta}$, whereas in the case of equation (7) we would like to estimate \mathbf{x}_p^o for a given $\hat{\mathbf{A}}$. The form of the solution is analogous to that for generalized least squares regression and yields

$$\hat{\mathbf{x}}_p = (\hat{\mathbf{A}}^T \boldsymbol{\Psi}^{-1} \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T \boldsymbol{\Psi}^{-1} \mathbf{x} \tag{9}$$

Here $\hat{\mathbf{x}}_p$ is the maximum likelihood estimate of \mathbf{x}_p^o . Substitution back into the model equation gives the maximum likelihood estimates for the remaining elements of \mathbf{x} :

$$\hat{\mathbf{x}} = \hat{\mathbf{A}} \hat{\mathbf{x}}_p = \hat{\mathbf{A}} (\hat{\mathbf{A}}^T \boldsymbol{\Psi}^{-1} \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T \boldsymbol{\Psi}^{-1} \mathbf{x} \tag{10}$$

A more formal derivation of equation (10) is given in Appendix II.

Equation (10) solves the first of the problems posed, allowing maximum likelihood estimates of the measurements to be obtained for a given set of model parameters. However, to obtain the maximum likelihood fit, it is necessary to adjust the model coefficients ($\hat{\mathbf{A}}$) to minimize the objective function S^2 . In the case of uncorrelated errors this objective function is given by equation (1). In the case where errors are correlated among the rows, a more general form of equation (1) is minimized

$$S^2 = \sum_{j=1}^n (\mathbf{x}_j - \hat{\mathbf{x}}_j)^T \boldsymbol{\Psi}_j^{-1} (\mathbf{x}_j - \hat{\mathbf{x}}_j) = \sum_{j=1}^n \Delta \mathbf{x}_j^T \boldsymbol{\Psi}_j^{-1} \Delta \mathbf{x}_j \tag{11}$$

where, as before, \mathbf{x}_j represents a column vector of \mathbf{X} . Equation (11) reduces to equation (1) for a diagonal covariance matrix. For the case where the error covariance matrix $\boldsymbol{\Psi}$ is the same for each column of \mathbf{X} , Fuller has given a closed-form solution for $\hat{\mathbf{A}}$ that minimizes S^2 (Reference 14, p. 292).

If Ψ is also diagonal, this is equivalent to the solution obtained by SVD if appropriate row scaling is used. However, in the general case where the error covariance matrix varies with the columns of \mathbf{X} , there is no closed-form solution for $\hat{\mathbf{A}}$. Fuller suggests an iterative solution in this case (Reference 14, p. 217). We have successfully employed both simplex and gradient-based algorithms to optimize the coefficients of $\hat{\mathbf{A}}$, but in general convergence is slow and prone to local minima. Furthermore, depending on which rows are used for the 'independent' variables, the numerical stability of the solution algorithm is questionable. Another drawback to this approach is that the equations developed thus far are in the form of a regression model rather than the PCA model which is sought. For these reasons it would be more convenient to represent equation (10) in terms of a PCA decomposition, i.e. in terms of scores and loadings. To do this, consider the form of the PCA model normally arrived at through SVD:

$$\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T \quad (12)$$

where $\hat{\mathbf{X}}$ is $m \times n$, $\hat{\mathbf{U}}$ is $m \times p$, $\hat{\mathbf{S}}$ is $p \times p$ and $\hat{\mathbf{V}}$ is $n \times p$. The caret on \mathbf{X} denotes that these are the maximum likelihood estimates of the measurements in accordance with the p -dimensional model and $\hat{\mathbf{U}}$, $\hat{\mathbf{S}}$ and $\hat{\mathbf{V}}$ are obtained from the singular value decomposition of $\hat{\mathbf{X}}$, which is constrained to be rank p . Now $\hat{\mathbf{X}}$ and $\hat{\mathbf{U}}$ are partitioned into the upper p rows ($\hat{\mathbf{X}}_1$ and $\hat{\mathbf{U}}_1$) and the lower $m-p$ rows ($\hat{\mathbf{X}}_2$ and $\hat{\mathbf{U}}_2$) to give

$$\hat{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{X}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{U}}_1 \\ \hat{\mathbf{U}}_2 \end{bmatrix} \hat{\mathbf{S}} \hat{\mathbf{V}}^T = \left(\begin{bmatrix} \hat{\mathbf{U}}_1 \\ \hat{\mathbf{U}}_2 \end{bmatrix} \hat{\mathbf{U}}_1^{-1} \right) \hat{\mathbf{U}}_1 \hat{\mathbf{S}} \hat{\mathbf{V}}^T = \begin{bmatrix} \mathbf{I}_p \\ \hat{\mathbf{U}}_2 \hat{\mathbf{U}}_1^{-1} \end{bmatrix} \hat{\mathbf{X}}_1 \quad (13)$$

or, with reference to equation (10),

$$\hat{\mathbf{A}} = \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \quad (14)$$

Thus there is a direct relationship between the parametric equations and SVD form of the model in the absence of intercepts. Substituting this into equation (10) yields

$$\begin{aligned} \hat{\mathbf{x}}_j &= \hat{\mathbf{A}}(\mathbf{A}^T \Psi_j^{-1} \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T \Psi_j^{-1} \mathbf{x}_j \\ &= \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} [(\hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1})^T \Psi_j^{-1} \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1}]^{-1} (\hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1})^T \Psi_j^{-1} \mathbf{x}_j \\ &= \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} [(\hat{\mathbf{U}}_1^{-1})^T \hat{\mathbf{U}}^T \Psi_j^{-1} \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1}]^{-1} (\hat{\mathbf{U}}_1^{-1})^T \hat{\mathbf{U}}^T \Psi_j^{-1} \mathbf{x}_j \\ &= \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \hat{\mathbf{U}}_1 (\hat{\mathbf{U}}^T \Psi_j^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}_1^T (\hat{\mathbf{U}}_1^{-1})^T \hat{\mathbf{U}}^T \Psi_j^{-1} \mathbf{x}_j \\ &= \hat{\mathbf{U}} (\hat{\mathbf{U}}^T \Psi_j^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^T \Psi_j^{-1} \mathbf{x}_j \\ &= \mathbf{P}_j \mathbf{x}_j \end{aligned} \quad (15)$$

where the projection matrix \mathbf{P}_j is given by

$$\mathbf{P}_j = \hat{\mathbf{U}} (\hat{\mathbf{U}}^T \Psi_j^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^T \Psi_j^{-1} \quad (16)$$

Like equation (10), equation (15) allows the maximum likelihood values for the matrix of measurements to be calculated, but in accordance with a given SVD model rather than a regression model. It is also similar to an equation developed by Bartlett in the psychometrics literature as early as 1937^{15,16} and discussed later by Lawley and Maxwell,⁸ who describe it as an unbiased method for estimating factor scores. It should be noted that because $\hat{\mathbf{S}}$ is diagonal the score matrix $\hat{\mathbf{R}}$ ($=\hat{\mathbf{U}}\hat{\mathbf{S}}$) can be substituted for $\hat{\mathbf{U}}$ in equations (14)–(16). In order for equation (15) to provide maximum likelihood estimates, the measurement errors should be normally distributed and the error covariance matrix

needs to be available. In the case of uncorrelated errors, Ψ will be a diagonal matrix with the diagonal elements equal to the measurement variances.

As before, equation (15) is used to optimize the elements of \hat{U} in accordance with the objective function given in equation (11). In this case, however, there should be fewer parameters to optimize. While the matrix \hat{A} has $p(m-p)$ variable coefficients, the columns of \hat{U} define an orthonormal set of vectors in the row space and it is only necessary to optimize $m-1$ angles in this space to define the optimum hyperplane.

To optimize the SVD model, an initial estimate for \hat{U} , designated \hat{U}_0 , is first obtained. The column vectors of \hat{U}_0 are then rotated in the m -dimensional space by applying an $m \times m$ rotation matrix T . This gives a new estimate for \hat{U} :

$$\hat{U} = T\hat{U}_0 \tag{17}$$

One easy way to define the rotation matrix is in terms of successive rotations about each axis. In an m -dimensional space there are $m-1$ rotation angles to be specified, so we have

$$T = T_1 T_2 \dots T_{m-1} = \prod_{i=1}^{m-1} T_i \tag{18}$$

where

$$T_1 = \begin{bmatrix} \cos \alpha_1 & -\sin \alpha_1 & 0 & \dots & 0 \\ \sin \alpha_1 & \cos \alpha_1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \cos \alpha_2 & -\sin \alpha_2 & \dots & 0 \\ 0 & \sin \alpha_2 & \cos \alpha_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \quad \text{etc.} \tag{19}$$

The problem now is one of optimizing the rotation angles $\alpha_1, \dots, \alpha_{m-1}$ to minimize the objective function in equation (11).

The optimization of the rotation angles can be carried out by a number of methods, but gradient methods are generally regarded as being the most efficient. These require the calculation of derivatives of S^2 with respect to the rotation angles. Since this is a nontrivial calculation, the derivation is included as Appendix III. The result is given as

$$\frac{\partial S^2}{\partial \alpha_i} = -2 \sum_{j=1}^n (\Delta x_j^T \Psi_j^{-1} G_i P_j x_j - x_j^T \Psi_j^{-1} P_j G_i P_j \Delta x_j - \Delta x_j^T \Psi_j^{-1} P_j G_i P_j x_j + x_j^T \Psi_j^{-1} G_i P_j \Delta x_j) \tag{20}$$

where the matrix G_i is defined in Appendix III. This equation has been checked against numerically calculated derivatives and found to be correct. It can be used in conjunction with standard gradient techniques to find the optimum rotation of eigenvectors to minimize the objective function in equation (11). In practice this procedure is faster and more reliable than using the regression form of the equation, but it is still relatively slow and susceptible to local minima. Therefore an alternative approach was sought. This is described in the next subsection.

3.2. An efficient MLPCA algorithm

In order to be useful with large data sets, an MLPCA procedure is needed which converges relatively quickly. Among the most efficient methods in this regard are iterative procedures, such as alternating regression approaches. Such a solution was developed for the MLPCA problem and was based on the

following rationale.

It will be assumed for the moment that all measurement errors are independent so that error covariance matrices in both the row and column spaces are diagonal. If this is true, the p -dimensional model obtained by maximum likelihood estimation must be equivalent in both spaces. This follows because the objective function in both cases reduces to the same summation given by equation (1). Mathematically,

$$S^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ij} - \hat{x}_{ij})^2}{\sigma_{ij}^2} = \sum_{j=1}^n \Delta \mathbf{x}_j^T \Psi_j^{-1} \Delta \mathbf{x}_j = \sum_{i=1}^m \Delta \mathbf{x}_i^T \Sigma_i^{-1} \Delta \mathbf{x}_i \quad (21)$$

Here $\Delta \mathbf{x}_j$ is a column vector of $\Delta \mathbf{X}$, $\Delta \mathbf{x}_i$ is a column vector of $(\Delta \mathbf{X})^T$, and Ψ_j and Σ_i are the corresponding column and row error covariance matrices for \mathbf{X} , both of which are diagonal. For ease of visualization, some of the matrices are represented pictorially in Figure 1(a). In order to develop the alternating regression algorithm, equation (12) will be rewritten as

$$\hat{\mathbf{X}}^T = \hat{\mathbf{V}} \hat{\mathbf{S}} \hat{\mathbf{U}}^T \quad (22)$$

This suggests that the maximum likelihood estimates of the measurements in the column space are given by an equation which is analogous to equation (15):

$$\hat{\mathbf{x}}_i = \hat{\mathbf{V}} (\hat{\mathbf{V}}^T \Sigma_i^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \Sigma_i^{-1} \mathbf{x}_i \quad (23)$$

where, as before, \mathbf{x}_i is a column vector of \mathbf{X}^T and Σ_i is the corresponding error covariance matrix. When the maximum likelihood solution has been obtained, the estimates of \mathbf{X} in the row and column spaces will be identical. This implies that an alternating regression approach can be developed by alternately transposing the maximum likelihood estimates and performing SVD.

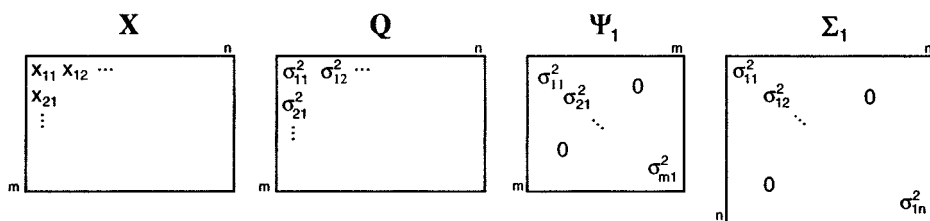
The algorithm for the alternating regression procedure is given in Table 1. It should be noted that the algorithm has been expanded to show a full iteration for clarity, but there are some redundancies in the procedure that can be exploited to make the actual code more compact. The algorithm alternately uses the maximum likelihood estimates in the original row space to update the estimates in the column space (i.e. the row space of the transposed matrix) and vice versa. This procedure has been found to be simple, fast and reliable. It does not appear to be susceptible to local minima, as is the case for gradient methods. Convergence time will depend on the dimensionality of the problem, the accuracy of the initial SVD estimate and the structure of the errors. The algorithm is easily applied to cases where there are missing data simply by incorporating large variances for the missing measurements. Convergence is somewhat slower in these cases, owing to the poor initial estimates obtained when the missing measurements are replaced with zeros, but is still reliable. Some comparative data on convergence times are given in Section 5.1.

The algorithm presented in Table 1 does impose certain restrictions. First, it is assumed that there are no offsets in the row or column space. Normally this would be equivalent to saying that the data have been mean centered but in the case of non-uniform measurement errors, mean centering is not generally equivalent to eliminating offsets. The topic of row and column offsets is discussed in Section 3.4.

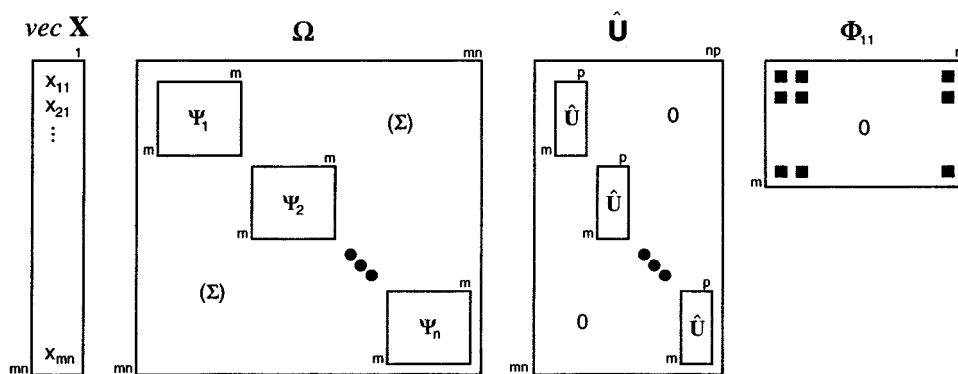
Another restriction to the algorithm presented here is that it assumes uncorrelated errors in the row and column spaces. The algorithm will not converge to a common solution if the covariance matrices used are not diagonal in both spaces. This raises an interesting question. Suppose, for example, that one is dealing with a series of m samples whose spectra are measured at n different wavelengths. Also

imagine that, because of instrumental characteristics, errors are correlated in the wavelength direction but there is no correlation in the errors among the samples. Under these conditions the $m \times m$ error covariance matrices in the row space, Ψ_j , are diagonal and minimization of S^2 should lead to the same solution regardless of whether the $n \times n$ error covariance matrices in the column space, Σ_i , are diagonal or not, since there is no information about wavelength correlation in the Ψ_j . However, it is apparent that the maximum likelihood estimates of \mathbf{X} obtained by equation (23) depend on whether or not the Σ_i are diagonal and will not be the same as the maximum likelihood estimates obtained by equation (15) if there are correlated errors. Therefore it seems that the maximum likelihood solution found in one space is not generally equivalent to that found in the alternate space in the presence of correlated errors. The reason for this apparent paradox is that the points in the row space are assumed to be independent, which will not be true if errors are correlated in the wavelength direction, so the model is invalid. The subject of correlated measurement errors is addressed in the next subsection.

(a) uncorrelated error



(b) correlated error



(c) intercept terms

$$\begin{matrix} m & & n \\ \boxed{\mathbf{B}} & = & \begin{matrix} 1 \\ \mathbf{1} \end{matrix} \begin{matrix} 1 & & n \\ \boxed{\mathbf{c}^T} \end{matrix} + \begin{matrix} 1 \\ \mathbf{d} \end{matrix} \begin{matrix} 1 & & n \\ \boxed{\mathbf{1}} \end{matrix} \\ m & & m \end{matrix}$$

Figure 1. Pictorial representation of some MLPCA matrices: (a) matrices used in standard algorithm; (b) matrices used in algorithm which incorporates error covariance; (c) composition of background matrix in algorithm incorporating intercept terms.

3.3. Error covariance

When measurements are made during the course of an experiment, there is a realistic possibility that random errors in these measurements will be correlated with one another because of the design of the experiment or the nature of the samples. Even if the original measurement errors are not correlated, it is possible that preprocessing methods such as digital filtering can introduce such correlation. To our knowledge, no one has attempted to develop PCA models which deal with correlated measurement errors, although there has been recognition of the importance of correlated errors in the literature.¹⁷ Earlier works cited^{6,7} attempt to develop algorithms to minimize equation (1), which assumes uncorrelated errors for maximum likelihood estimation. In the more general case we wish to minimize equation (11), which incorporates the non-diagonal error covariance matrix. Furthermore, the model developed should be consistent with an SVD formulation such that the maximum likelihood estimates obtained in either the row or the column space will be the same. In practice one is fortunate to have individual measurement standard deviations available, while information on error covariance is rare.

Table 1. Standard MLPCA algorithm (uncorrelated errors, no intercepts)

1. Given an $m \times n$ data matrix \mathbf{X} and a corresponding $m \times n$ matrix \mathbf{Q} of measurement error variances, use SVD to obtain an initial approximation to the MLPCA solution. The SVD solution is truncated to rank p as indicated by the notation $\text{svd}(\mathbf{X}, p)$. This means that \mathbf{U} , \mathbf{S} and \mathbf{V} are truncated to $m \times p$, $p \times p$ and $n \times p$ respectively.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\mathbf{X}, p) \quad (\text{T1})$$

2. Transpose \mathbf{X} and \mathbf{Q} and calculate the maximum likelihood estimates in the alternate space using $\hat{\mathbf{V}}$.

$$\mathbf{X} \leftarrow \mathbf{X}^T, \quad \mathbf{Q} \leftarrow \mathbf{Q}^T, \quad \boldsymbol{\Sigma}_i \leftarrow \text{diag}(\mathbf{q}_i) \quad (\text{T2})$$

$$\hat{\mathbf{x}}_i = \hat{\mathbf{V}}(\hat{\mathbf{V}}^T \boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{x}_i \quad (\text{T3})$$

Here \mathbf{x}_i is a column vector of the now transposed \mathbf{X} . From this result the objective function can be calculated using equation (T4).

$$S_1^2 = \sum_{i=1}^m (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) = \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ji} - \hat{x}_{ji})^2}{\sigma_{ji}^2} \quad (\text{T4})$$

3. Compute the SVD of $\hat{\mathbf{X}}$ from step 2 and, as before, truncate the results to obtain a new $\hat{\mathbf{V}}$.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\hat{\mathbf{X}}, p) \quad (\text{T5})$$

4. Repeat step 2 to estimate the model in the original space.

$$\mathbf{X} \leftarrow \mathbf{X}^T, \quad \mathbf{Q} \leftarrow \mathbf{Q}^T, \quad \boldsymbol{\Psi}_j \leftarrow \text{diag}(\mathbf{q}_j) \quad (\text{T6})$$

$$\hat{\mathbf{x}}_j = \hat{\mathbf{V}}(\hat{\mathbf{V}}^T \boldsymbol{\Psi}_j^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \boldsymbol{\Psi}_j^{-1} \mathbf{x}_j \quad (\text{T7})$$

$$S_2^2 = \sum_{j=1}^n (\mathbf{x}_j - \hat{\mathbf{x}}_j)^T \boldsymbol{\Psi}_j^{-1} (\mathbf{x}_j - \hat{\mathbf{x}}_j) = \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ij} - \hat{x}_{ij})^2}{\sigma_{ij}^2} \quad (\text{T8})$$

5. Compute the SVD of $\hat{\mathbf{X}}$ to obtain a new estimate of the MLPCA solution in the original space.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\hat{\mathbf{X}}, p) \quad (\text{T9})$$

6. Calculate the convergence parameter λ .

$$\lambda = (S_1^2 - S_2^2) / S_2^2 \quad (\text{T10})$$

If λ is less than the convergence limit (typically 10^{-10} in this work), terminate. Otherwise return to step 2.

Nevertheless, it is useful to develop a theoretical framework for using such information if for no other reason than to assess its value.

There are essentially three common cases of error correlation that can be distinguished: (i) all measurement errors are uncorrelated; (ii) correlations among errors exist along either the rows or columns of the data matrix, but the errors are uncorrelated in the other direction; (iii) there is some degree of possible correlation among all the measurement errors. The first case was dealt with in the preceding two subsections and the third is the completely general case which has yet to be addressed. To begin, however, it is helpful to examine the second case, which is more restricted.

An example of the second case was presented earlier and will be considered again here. Consider a series of spectra whose errors are correlated in the wavelength direction (e.g. by source fluctuations) but not correlated among samples. If this were true, the error covariance matrix for column j of \mathbf{X} (m samples by n wavelengths), Ψ_j , would be diagonal, but that for row i , Σ_i , would not. Variance information is carried in both spaces, but covariance information is only carried in the column space in this case. Therefore it would seem logical to compute the maximum likelihood estimates using equation (23) and minimize the objective function by rotating the columns of \mathbf{V}_0 . In principle this can be done and will lead to the correct result. However, it would have to be done using the gradient methods described in Section 3.1 rather than the much more efficient algorithm described in Section 3.2, since we can no longer interchange the row and column spaces. Furthermore, when the final solution is obtained, the maximum likelihood estimates of \mathbf{X} computed by equation (23) in the column space will not be the same as those calculated in the row space using equation (15). This is an apparent contradiction, since there should only be one set of maximum likelihood projections which are the same in either space. The reason for this paradox is that there is no information about wavelength correlation in the row space, so the maximum likelihood estimates generated there are wrong. Realizing this, one could simply use the estimates obtained from equation (23), but this does not address the more general problem of incorporating the error covariance information in both spaces.

To arrive at a more general solution for correlated errors, it is necessary to realize that any pair of measurement errors could be correlated and redefine the problem accordingly. Rather than considering it as modeling n points in an m -dimensional space or m points in an n -dimensional space, it will be viewed as modeling a single point in an mn -dimensional space. To do this, \mathbf{X} is vectorized by applying the 'vec' operator and the equations are adapted as necessary. The generalizations of equations (15) and (11) are

$$\text{vec}(\hat{\mathbf{X}}) = \hat{\mathbf{U}}(\hat{\mathbf{U}}^T \mathbf{\Omega}^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^T \mathbf{\Omega}^{-1} \text{vec}(\mathbf{X}) \tag{24}$$

$$S^2 = \text{vec}(\Delta \mathbf{X})^T \mathbf{\Omega}^{-1} \text{vec}(\Delta \mathbf{X}) \tag{25}$$

where

$$\hat{\mathbf{U}} = \mathbf{I}_n \otimes \hat{\mathbf{U}} \tag{26}$$

$$\mathbf{\Omega} = E[(\text{vec}(\mathbf{X} - \mathbf{X}^0)) \cdot (\text{vec}(\mathbf{X} - \mathbf{X}^0))^T] \tag{27}$$

Here the 'vec' operator gives an $mn \times 1$ vector with the column vectors of \mathbf{X} arranged in sequence.¹⁸ The symbol ' \otimes ' indicates the Kronecker product such that each element of \mathbf{I}_n is multiplied by $\hat{\mathbf{U}}$.¹⁸ Thus $\hat{\mathbf{U}}$ is an $mn \times np$ matrix with $\hat{\mathbf{U}}$ ($m \times p$) repeating along the diagonal. $\mathbf{\Omega}$ is the full covariance matrix for $\text{vec}(\mathbf{X})$, providing the error covariance among all the measurements. \mathbf{X}^0 represents the true (or expectation) values for \mathbf{X} . For greater clarity, some of these matrices are shown pictorially in Figure 1(b). Note that the column variance matrices of \mathbf{X} , represented as Ψ , fall along the diagonal of $\mathbf{\Omega}$. The remainder of $\mathbf{\Omega}$ is made up with the row covariance information (Σ) and other covariances.

With these definitions an alternating regression algorithm similar to the one in the preceding

subsection can be developed and is given in Table 2. As before, the algorithm above uses the maximum likelihood estimates in one space to estimate the solution in the alternate space. As the solutions are exchanged, the error covariance matrix for $\text{vec}(\mathbf{X})$ (given by $\mathbf{\Omega}$) needs to be modified to give the covariance matrix for $\text{vec}(\mathbf{X}^T)$ (given by $\mathbf{\Xi}$). This can be done on an element-by-element basis, but it is easier to use the commutation matrix \mathbf{K} .¹⁸ The commutation matrix is an orthonormal matrix that has the property

$$\text{vec}(\mathbf{A}^T) = \mathbf{K} \text{vec}(\mathbf{A}) \quad (28)$$

When combined with the definition in equation (25), this leads to the use of equation (55) (Appendix III) to transform the error covariance matrix into the alternate space. In practice the commutation matrix can be computed as follows. Begin with an $mn \times 1$ vector \mathbf{a} such that $a_i = i$. Reshape \mathbf{a} so that it forms the $m \times n$ matrix \mathbf{A} and then set $\mathbf{b} = \text{vec}(\mathbf{A}^T)$. Now the corresponding elements of \mathbf{a} and \mathbf{b} are the row and column indices respectively of the elements of the $mn \times mn$ commutation matrix \mathbf{K} that should be set to one. The remaining elements of \mathbf{K} should be set to zero, making it a sparse matrix with mn non-zero elements.

The algorithm in Table 2 represents a completely general treatment for the case of correlated measurement errors and therefore is a significant advance in multivariate modeling. It converges rapidly to an optimal solution (unless the matrices involved are numerically unstable) and yields

Table 2. MLPCA algorithm for correlated measurement errors

1. Given an $m \times n$ data matrix \mathbf{X} , a corresponding $mn \times mn$ matrix $\mathbf{\Omega}$ of measurement error covariances for $\text{vec}(\mathbf{X})$ and a commutation matrix \mathbf{K} for \mathbf{X} , use a truncated SVD to obtain an initial approximation to the MLPCA solution.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\mathbf{X}, p) \quad (T11)$$

2. Transpose \mathbf{X} and calculate the maximum likelihood estimates in the alternate space using $\hat{\mathbf{V}}$.

$$\mathbf{X} \leftarrow \mathbf{X}^T, \quad \mathbf{\Xi}^{-1} \leftarrow \mathbf{K} \mathbf{\Omega}^{-1} \mathbf{K}^T \quad (T12)$$

$$\hat{\mathbf{V}} = \mathbf{I}_m \otimes \hat{\mathbf{V}} \quad (T13)$$

$$\text{vec}(\hat{\mathbf{X}}) = \hat{\mathbf{V}} (\hat{\mathbf{V}}^T \mathbf{\Xi}^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \mathbf{\Xi}^{-1} \text{vec}(\mathbf{X}) \quad (T14)$$

From this result the objective function can be calculated using equation (T16).

$$S_1^2 = \text{vec}(\Delta \mathbf{X})^T \mathbf{\Xi}^{-1} \text{vec}(\Delta \mathbf{X}) \quad (T15)$$

3. Reconstruct $\hat{\mathbf{X}}$ from $\text{vec}(\hat{\mathbf{X}})$ and compute the truncated SVD of $\hat{\mathbf{X}}$.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\hat{\mathbf{X}}, p) \quad (T16)$$

4. Repeat step 2 to estimate the model in the original space.

$$\mathbf{X} \leftarrow \mathbf{X}^T, \quad \mathbf{\Omega}^{-1} \leftarrow \mathbf{K}^T \mathbf{\Xi}^{-1} \mathbf{K} \quad (T17)$$

$$\hat{\mathbf{V}} = \mathbf{I}_n \otimes \hat{\mathbf{V}} \quad (T18)$$

$$\text{vec}(\hat{\mathbf{X}}) = \hat{\mathbf{V}} (\hat{\mathbf{V}}^T \mathbf{\Omega}^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \mathbf{\Omega}^{-1} \text{vec}(\mathbf{X}) \quad (T19)$$

$$S_2^2 = \text{vec}(\Delta \mathbf{X})^T \mathbf{\Omega}^{-1} \text{vec}(\Delta \mathbf{X}) \quad (T20)$$

5. Reconstruct $\hat{\mathbf{X}}$ (original dimensions) and compute the truncated SVD of $\hat{\mathbf{X}}$ in the original space.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\hat{\mathbf{X}}, p) \quad (T21)$$

6. Compute the convergence parameter (equation (T10)) and terminate if it is less than the convergence limit. Otherwise return to step 2.

results identical to the earlier algorithm in the presence of uncorrelated errors. In practice, use of the algorithm is currently limited to some extent by the size and stability of the matrices. In the completely general case the covariance matrix will have m^2n^2 elements and easily exceeds the storage capacity of most machines for large matrices unless special measures are used. The matrices also tend to become ill-conditioned as \mathbf{X} becomes large, causing convergence problems. However, for many chemical problems, error covariance is limited to either the row or column directions. In these cases either $\mathbf{\Omega}$ or $\mathbf{\Xi}$ will be block diagonal and can be stored as a sparse matrix. The diagonal blocks of these matrices ($\mathbf{\Psi}$ or $\mathbf{\Sigma}$) can be inverted individually and the covariance matrix in the alternate space can be calculated with the commutation matrix. In this way the algorithm can be extended to a much wider set of problems.

3.4. MLPCA with intercepts

In models for chemical systems it is common for row and column offsets to be present for the matrix of measurements. Returning to the earlier example from spectroscopy, one can imagine a situation in which a constant background spectrum is present for all the samples. If \mathbf{X} is m samples by n wavelengths, this can be considered a vector of column offsets. Since this is invariant for all samples, it is often desirable to subtract it from each sample spectrum prior to decomposition of the data matrix, thereby achieving a reduction rank. Likewise, one can imagine a vector of row offsets that arises from, say, variations in cell position or sample preparation. This can also be removed. Models which include such effects in chemistry are the same as that developed by Mandel for analysis of variance,¹⁹ namely

$$x_{ij} = \mu + \rho_i + \gamma_j + \sum_{k=1}^p u_{ik} s_{kk} v_{jk} \quad (29)$$

Here μ is the grand mean of \mathbf{X} , ρ_i and γ_j represent row and column offsets respectively and u , s and v are individual elements of the SVD of the matrix with the offsets removed. The elements of the vectors $\mathbf{\rho}$ and $\mathbf{\gamma}$ are often taken to be the means of the rows and columns after the grand mean is subtracted. Note that equation (29) is a general formulation. In a given application the row and/or column offsets could be set to zero (as they often are) or could even be constrained, for example, to each have identical elements (a situation rarely imposed in chemistry). Also note that the grand mean, since it is constant, can be incorporated into either or both of the offset terms and so can be excluded from equation (29). For the purposes of this discussion an alternate form of equation (29) will be used:

$$\hat{\mathbf{X}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T + \mathbf{1}_m\mathbf{c}^T + \mathbf{d}\mathbf{1}_n^T = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T + \mathbf{B} \quad (30)$$

In this equation, \mathbf{c} and \mathbf{d} are column vectors of the row and column offsets respectively and $\mathbf{1}_m$ and $\mathbf{1}_n$ are column vectors of ones of length m and n . This representation of the matrix of offsets, \mathbf{B} , is shown pictorially in Figure 1(c). It is clear from the figure that the presence of row or column offsets will increase the rank of an untreated data matrix by one, while the presence of both will increase the rank by two.

In chemistry, realization of a model of the form of equation (30) is normally accomplished by column and/or row mean centering to determine \mathbf{c} and \mathbf{d} . If only one of these is to be used, the means of the columns or rows can be used directly as \mathbf{c} or \mathbf{d} . If both are used, the grand mean must first be subtracted from the data matrix or one set of offsets needs to be calculated after the other has been subtracted from \mathbf{X} .

It should be pointed out that there are an infinite number of row and column offset vectors which

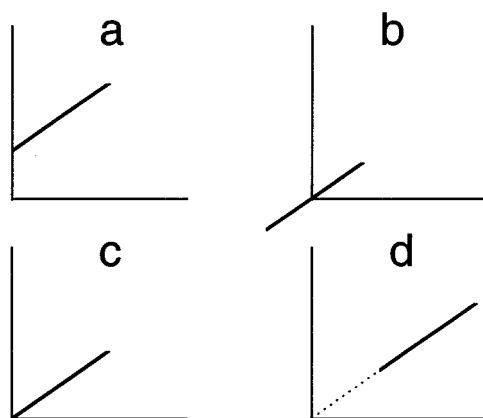


Figure 2. Representation of equivalent translations for removing model intercepts: (a) original data; (b) mean centered in x and y ; (c) offset of zero in x ; (d) offset of zero in y .

will provide equivalent models in terms of quality of fit. This is illustrated in Figure 2 for the case of rank-one data (with an offset) in a two-dimensional space. Note that a zero intercept for the model can be obtained in at least three ways and there are infinitely more. In fact, the illustration shows that any one of the offset terms (x or y) can be set to zero without changing the quality of the fit. In general, any p offset terms can be set to zero for both **c** and **d**, which is why the degrees of freedom are reduced by $n - p$ and/or $m - p$ when mean centering is used.

When all the measurements in \mathbf{X} have normal i.i.d. errors, mean centering to remove offsets is a convenient approach to use because the characteristics of PCA *guarantee* that the mean will fall on the optimum model, so forcing the mean to zero ensures that all the intercept terms will also be zero for the centered data. However, for MLPCA the presence of non-uniform and/or correlated error distributions means that this is no longer generally true, although it may be a good approximation. For this reason the row and column offset vectors need to be optimized along with the scores and loadings in order to obtain a true maximum likelihood solution. Attempts to include these parameters in the alternating regression algorithms already presented have not been successful thus far and generally result in convergence to a suboptimal solution. As an alternative, more traditional gradient methods have been coupled with the alternating regression procedure to yield the MLPCA solution. Although this is slower than the standard MLPCA algorithm, it converges reliably.

The procedure begins by finding initial estimates for the row and/or column offset vectors. One way to do this would be to use the corresponding means, but we chose an alternate procedure. The data are first analyzed using the algorithm with no intercepts, but increasing the model rank by one or two to account for the offset vectors. The row and/or column means of the maximum likelihood estimates are then used as a starting point for the offset parameters. As noted above, the procedure should require the optimization of only $n - p$ row offsets and/or $m - p$ column offsets. However, the full vectors in each direction (n and/or m parameters) have been used in this work to simplify the conversion from and comparison with the row and column means. This will lead to degenerate solutions but does not seem to affect convergence.

Once initial estimates for **c** and **d** have been obtained, these are used to calculate **B** (see equation (30)) and this is subtracted from \mathbf{X} . The alternating regression algorithm is then applied to the adjusted matrix. As soon as the convergence criterion has fallen below an acceptable value, a gradient search is implemented to optimize **c** and/or **d**. The results of this are used to calculate a new **B**, which is then subtracted from the original \mathbf{X} , and the process is repeated until the change in the objective function

is acceptably small. In order to carry out the gradient optimization, the derivative of S^2 with respect to the intercept parameters is needed. For uncorrelated errors this can be obtained by using the equation for $\Delta \mathbf{x}$ in the presence of intercepts:

$$\Delta \mathbf{x}_j = [\mathbf{I} - \hat{\mathbf{U}}(\hat{\mathbf{U}}^T \boldsymbol{\Psi}_j^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^T \boldsymbol{\Psi}_j^{-1}] (\mathbf{x}_j - \mathbf{b}_j) \quad (31)$$

Here \mathbf{b}_j is a column vector of \mathbf{B} . This gives

$$\frac{\partial S^2}{\partial c_j} = \Delta \mathbf{x}_j^T \boldsymbol{\Psi}_j^{-1} [\mathbf{I} - \hat{\mathbf{U}}(\hat{\mathbf{U}}^T \boldsymbol{\Psi}_j^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^T \boldsymbol{\Psi}_j^{-1}] \mathbf{1}_m \quad (32)$$

$$\frac{\partial S^2}{\partial \mathbf{d}} = \sum_{j=1}^n \{ \Delta \mathbf{x}_j^T \boldsymbol{\Psi}_j^{-1} [\mathbf{I} - \hat{\mathbf{U}}(\hat{\mathbf{U}}^T \boldsymbol{\Psi}_j^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^T \boldsymbol{\Psi}_j^{-1}] \}^T \quad (33)$$

These equations were employed for the gradient optimization.

4. EXPERIMENTAL

All the calculations performed in this work were carried out on a DEC 3000/300X Unix-based workstation with a clock speed of 175 MHz and 96 MB of memory (Digital Equipment Corp., Maynard, MA). Programs were written in Matlab v. 4.2a (The Maths Works Inc., Natick, MA). Seven data sets were used to evaluate the algorithms.

Data set 1 was a simulated rank-two data set of dimensions 10×20 . The error-free data matrix was generated by multiplying a 10×2 matrix of elements from a uniform distribution of random numbers between zero and one ($U(0, 1)$) by a 2×20 matrix that was also drawn from $U(0, 1)$. Measurement standard deviations corresponding to this 10×20 matrix were determined by generating a 10×20 matrix of random numbers from $U(0, 0.01)$. This ensured that there was no pattern in the standard deviations. Finally, the 10×20 matrix of measurement errors was generated by taking 10×20 matrix of normally distributed random numbers (mean=0, standard deviation=1, or $N(0, 1)$) and multiplying this on an element-by-element basis by the matrix of standard deviations. The result was added to the error-free matrix to give the noisy data matrix \mathbf{X} . The matrix of variances, \mathbf{Q} , was obtained by squaring the elements of the standard deviation matrix. The matrices \mathbf{X} and \mathbf{Q} were passed to the MLPCA algorithm for uncorrelated errors.

Data sets 2 and 3 were also rank-two matrices generated in the same manner as data set 1, except that their dimensions were 20×20 and 20×100 .

Data set 4 was simulated rank-three spectral data. Pure component spectra were simulated as three Gaussian profiles spaced 20 nm apart, each with a standard deviation of 20 nm and a maximum height of unity. The maximum of the center profile was at 500 nm and 41 equally spaced points were calculated for each spectrum in the range 400–600 nm. A 20×3 concentration matrix was generated by drawing random numbers from a $U(0, 1)$ distribution. The 20×41 error-free matrix was the product of the concentration matrix and the 3×41 matrix of spectral profiles. To provide a matrix of standard deviations that was unstructured (i.e. rank greater than one) but still realistic, constant and proportional errors were used. The constant part was taken to be 1% of the maximum value of the noise-free data matrix. The 20×41 matrix of proportional standard deviations was calculated as 5% of the elements in the error-free data matrix. The overall matrix of standard deviations was the square root of the sum of the squares of the proportional part and the constant part. Finally, random numbers from an $N(0, 1)$ distribution were multiplied by each element of the standard deviation matrix to give the error matrix, which was added to the error-free data to give the noisy data matrix \mathbf{X} .

Data set 5 was generated in exactly the same manner as data set 4, except that random offsets drawn from an $N(0, 0.1)$ distribution were added to each row and column of the final \mathbf{X} . This was intended to test the version of the MLPCA algorithm designed to fit intercept terms.

Data set 6 consisted of near-infrared spectroscopic data for three-component mixtures containing toluene, chlorobenzene and heptane and was part of the Infomertix (Seattle, WA) calibration transfer study.²⁰ The 31 samples derived from an augmented, three-level, three-factor, full factorial design. The concentration was varied between 20% and 70% by weight for toluene and chlorobenzene and between 2% and 10% by weight for heptane. The mixtures were sealed into standard 1 cm path length cuvettes. Spectra were obtained over the range 400–2500 nm on an NIRSystems Model 6500 (NIRSystems, Silver Spring, MD) grating spectrometer at intervals of 2 nm and were the average result of 32 scans. The spectrometer employed a Si detector in the range 400–1100 nm and a PbS detector at longer wavelengths. A typical spectrum is shown in Figure 3(a). Clearly there are some regions above 1600 nm which are essentially opaque and are not meaningful, but these were retained in this study for the purpose of illustrating the features of MLPCA. Unfortunately, replicate data were only available for the first sample, for which 400 spectra had been obtained, so a complete matrix of standard deviations could not be constructed. Instead, the standard deviation data for the first sample, shown in Figure 3(b), was used for all the samples. Although this is not completely accurate, it should serve as a reasonable approximation, especially for regions where the standard deviation is very large owing to high absorbance.

Data set 7 was a 5×10 matrix constructed in the same manner as a data set 1, except that correlated errors were introduced. To produce error covariance, a 3×3 moving average filter (coefficients = 1/9) was applied to the 5×10 matrix of errors before it was added to the error-free measurements. At the boundaries of the error matrix the filter was wrapped around to the opposite side in order to eliminate

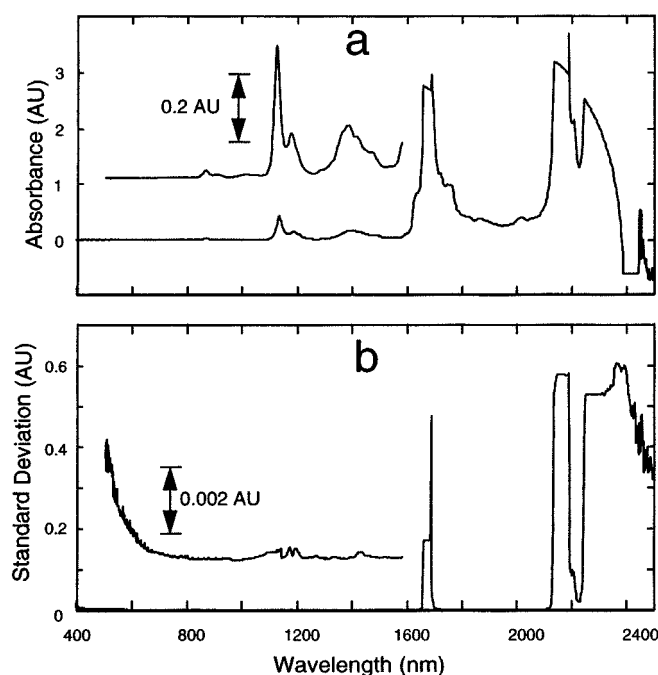


Figure 3. (a) Typical near-infrared spectrum of three-component mixture. (b) Standard deviation of absorbance measurements.

edge effects. Although this approach is not particularly realistic, it represents a general case for which the covariance structure could be easily predicted. The covariance matrix for this data set was calculated using the definitions

$$\mathbf{F} = [\text{vec}(\Phi_{11}) \text{vec}(\Phi_{21}) \dots \text{vec}(\Phi_{mm})] \quad (34)$$

$$\mathbf{\Omega} = \mathbf{F}^T \text{diag}(\text{vec}(\mathbf{Q}))\mathbf{F} \quad (35)$$

Here \mathbf{Q} is the 5×10 matrix of error variances prior to the application of the moving average filter. The 5×10 matrix Φ_{ij} contains the nine filter coefficients applied to the error matrix to give the filtered error corresponding to measurement ij . This is illustrated in Figure 1(b) for Φ_{11} , where the filled squares show the positions of the filter coefficients. For Φ_{12} the squares shift right and for Φ_{21} they shift down. Expressed another way, if \mathbf{E} represents the 5×10 matrix of uncorrelated errors generated in accordance with the variances in \mathbf{Q} and if ε_{ij} represents the error added to element ij of the pure data matrix, we have

$$\varepsilon_{ij} = [\text{vec}(\Phi_{ij})]^T \text{vec}(\mathbf{E}) \quad (36)$$

The errors generated in this way were added to the pure data matrix. The noisy data matrix \mathbf{X} and the error covariance matrix $\mathbf{\Omega}$ were passed to the MLPCA algorithm.

5. RESULTS AND DISCUSSION

5.1. Algorithm performance

Table 3 summarizes the results of applying the various MLPCA algorithms to the data sets described in the preceding section. For data sets 1–4 and 6 the standard algorithm (Table 1) was used. The version which incorporates intercept terms (Section 3.4) was used for data set 5 and the routine for correlated errors (Table 2) was employed for data set 7. In all cases the MLPCA model rank was varied from one to $p+1$, where p is the true rank of the data set. All the data sets were also analyzed by PCA, with mean centering used as a pretreatment step for data set 5. The models generated by PCA and MLPCA were used to estimate the error-free measurements ($\hat{\mathbf{X}}$) by orthogonal and maximum likelihood projections of the measurements respectively. These were used to calculate the objective function S^2 in each instance. For this purpose, equation (1) was used for data sets 1–6 and equation (25) was used for data set 7. For cases with no intercept terms, S^2 for the model with correct rank should approximate a χ^2 -distribution with $(m-p)(n-p)$ degrees of freedom. In accordance with this, the last two columns of Table 3 give the probability of realizing a value of S^2 below that observed if the model were correct. In other words, values of P below 0.025 or above 0.975 would constitute rejection of the null hypothesis that the model is correct ($\alpha=0.05$). For data set 5 the same test was done using $(m-p-1)(n-p-1)$ degrees of freedom to account for row and column intercepts.

The convergence times given in Table 3 are the result of single runs that were carried out with no competing tasks running on the computer. The results are 'typical' in the sense that no attempt was made to adjust the random number seeds to improve performance. Note that there is a separate convergence time listed for each rank analyzed by MLPCA. This is because, unlike conventional PCA, the MLPCA solutions are not generally nested (i.e. the rank- p model does not contain the rank- $(p-1)$ model). Convergence times listed are generally reasonable, with most cases requiring no more than a few minutes and all but one case requiring less than 1 h. Although size seems to play some role in convergence time, a much more important factor is the structure of the errors. Experience has indicated that the totally random error structure, such as that in data sets 1–3, is the most difficult case and this observation is supported by the results in Table 3. This case is therefore useful in estimating

Table 3. Results of MLPCA on test data sets

Data set number	Size	Error type	Offset?	True rank	Model rank	Convergence time (min)	S^2		P^a	
							PCA	MLPCA	PCA	MLPCA
1	10×20	Random	N	2	1	3.00	3.90×10^6	4.53×10^5	1.00	1.00
					2	1.46	3.29×10^4	123.04	1.00	0.10
					3	2.12	5209.0	88.48	1.00	0.02
2	20×20	Random	N	2	1	1.21	7.17×10^6	4.07×10^5	1.00	1.00
					2	1.65	1.51×10^4	311.82	1.00	0.32
					3	2.56	1.86×10^4	267.29	1.00	0.18
3	20×100	Random	N	2	1	700.50	8.71×10^9	7.13×10^6	1.00	1.00
					2	48.46	2.83×10^7	1757.3	1.00	0.46
					3	40.10	2.67×10^8	1548.3	1.00	0.04
4	20×41	Proportional + constant	N	3	1	0.03	8419.2	7100.9	1.00	1.00
					2	0.03	1240.2	1217.0	1.00	1.00
					3	0.03	724.90	683.43	0.98	0.85
					4	0.24	654.99	571.85	0.96	0.28
5	20×41	Proportional + constant	Y	3	1	13.73	8138.4	6592.5	1.00	1.00
					2	5.35	1094.4	953.02	1.00	1.00
					3	4.21	677.40	622.28	0.99	0.81
					4	12.78	603.87	512.46	0.97	0.20
6	31×1050	?	N	3	1	3.35	1.72×10^8	8.88×10^7	1.00	1.00
					2	8.13	1.63×10^8	1.44×10^7	1.00	1.00
					3	5.45	1.27×10^8	3.38×10^5	1.00	1.00
					4	10.84	5.39×10^7	1.43×10^5	1.00	1.00
7	5×10	Random, correlated	N	2	1	0.02	2.59×10^6	1.53×10^6	1.00	1.00
					2	0.11	199.92	23.01	1.00	0.48
					3	0.06	129.45	9.97	1.00	0.24

^a P is the probability that a value less than the corresponding S^2 would be observed for a χ^2 -distribution with the appropriate degrees of freedom.

upper limits for the convergence time. The rank-one model for data set 3 (the largest of the 'random' error models) proved to be unusually difficult to solve, requiring nearly 12 h. It is not clear at this point whether the slow convergence is the result of the error structure itself or poor initial estimates that arise from it. However, convergence seems to be considerably faster for other error structures, such as data set 4, where the convergence time is typically a few seconds. This case is likely to be much more typical of experimental data than the random structure. Data set 5, which is the same as data set 4 except for the presence of row and column offsets, required considerably more time because of the gradient optimization of intercept vectors as described in Section 3.4. Data set 6 represents a typical experimental data set and demonstrates that MLPCA is a practical alternative to PCA for such cases. Relatively slow convergence in this instance was probably due to poor initial estimates resulting from the inclusion of very noisy measurements in the data set. Finally, data set 7 shows that convergence time is not a problem for correlated errors.

Analysis of the objective function values in Table 3 shows that MLPCA always produces a lower S^2 than the corresponding PCA model. This is expected, since PCA does not optimize the same criterion. It is interesting to note, however, that for data sets 2 and 3 the value of S^2 obtained by PCA actually *increases* in going from a rank-two to a rank-three model. At first this may seem contradictory, since the general rule is that increasing the rank of a model improves the fit to the data. However, accounting for a greater amount of variance in the data set by increasing the number of factors in PCA does not necessarily decrease the value of S^2 . This seems to be especially true for the case of unstructured errors.

With reference to the last two columns of Table 3 it is expected that P should drop significantly below unity when a model of the correct rank is found, since that is the point at which S^2 should follow a χ^2 -distribution. For MLPCA this is true for all cases but one. The one exception is the experimental data, data set 6. In this case it is not surprising that P remains at unity for several reasons: (a) the matrix of standard deviations was approximated using only information from the first sample and therefore is incorrect for the remaining samples; (b) the variance estimates are the results of replicate scans and do not account for other sources of variance, such as cell positioning; (c) the variance estimates appear to reflect truncation of the signal in some places; (d) the noise is known to be correlated; (e) although there are only three known components in the system, there is a real possibility of row and/or column offsets. This example highlights some of the difficulties in using this sensitive statistical measure to estimate the rank in practical cases but does not diminish the utility of MLPCA for model estimation. (Note that the objective function for the rank-three model is more than two orders of magnitude smaller for MLPCA.) In contrast with MLPCA, the PCA models almost always give P -values of unity in the table, indicating an incorrect model even when the rank is overestimated. The only exceptions are data sets 4 and 5, where the error structure is closer to uniform, but even in those cases P does not fall below 0.98 for models of the correct rank.

5.2. Statistical validation

Although the cases in the preceding subsection indicate that MLPCA produces smaller values of the objective function than does PCA, it does not guarantee that the procedure converges on the optimum solution, since local optima are always possible. One way to test for a global optimum is to use a different initial estimate and compare the final solutions, but this method is not foolproof. A better method in this case is to exploit the statistical characteristics of S^2 for the correct model. This is done by analyzing replicate data sets, each with the same matrix of error-free data and standard deviations but with different errors. If the distribution of the S^2 -values for these replicates follows a χ^2 -distribution with the appropriate degrees of freedom, then it can be concluded that the method is finding the maximum likelihood solution.

A convenient way to make this comparison is to use probability plots. First the replicate data sets (100 in this case) are analyzed and the S^2 -values are stored. The S^2 -values are then sorted and assigned a cumulative probability according to their position in the list (the observed probability). For example, the second element in the list would be assigned an observed probability of $2/n$, where n is the number of replicates. Then an expected probability is calculated using the χ^2 -distribution. The cumulative probability density function for χ^2 can be calculated using the incomplete gamma function:²¹

$$P(S^2 | \nu) = \Gamma_{\text{inc}}\left(\frac{\nu}{2}, \frac{S^2}{2}\right) \quad (37)$$

where ν is the number of degrees of freedom. If the two distributions are the same, a plot of the expected probabilities against observed probabilities should yield a straight line with a slope of unity. If the model is insufficient to account for the systematic variance, either because the form of the model is incorrect or the parameters are suboptimal, then the points of the plot will lie above the ideal line. This means that the distribution of S^2 is shifted right from the χ^2 -distribution. If the model accounts for an excessive amount of variance (i.e. the estimated rank is too high and measurement variance is modeled), the points will lie below the ideal line.

Figure 4 shows probability plots for four of the data sets used in this study: data sets 2, 4, 5 and 7. These data sets were chosen to reflect the different error structures and algorithms used. It is clear from the figure that in all cases the results from MLPCA follow the expected distribution, with only minor

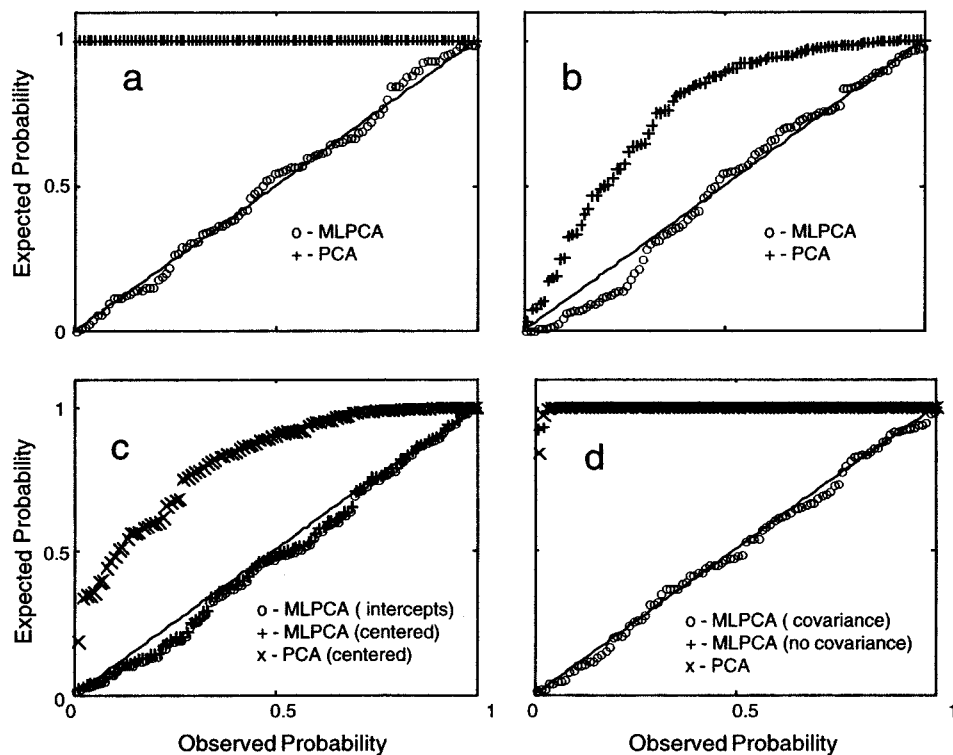


Figure 4. Probability plots of S^2 for replicate runs of simulated data sets: (a) data set 2; (b) data set 4; (c) data set 5; (d) data set 7.

deviations attributable to the statistical limitations of this study. Therefore it can be concluded that each of the algorithms is converging on a global optimum (the quality of the optimum is discussed in Section 5.5). Furthermore, to varying degrees, the models generated by PCA do not adequately account for the systematic variance in the data sets and are therefore likely to be inferior. Additional comments on these plots are made in the following two subsections.

5.3. MLPCA with intercepts

The results shown in Figure 4(c) confirm that the gradient optimization method used to determine the intercept vectors for MLPCA performed as expected. However, an additional analysis was done using the standard MLPCA algorithm after the data were column and row mean centered. While this approach always produced S^2 -values which were higher than those for which the intercepts were optimized, the difference is barely distinguishable in the figure. This implies that the intercepts determined are very close to the mean values for this particular error structure. Therefore, although the use of mean centering may mean that the models obtained by the standard MLPCA algorithm are suboptimal, the differences may be negligibly small in many cases and justify the use of the faster algorithm. Larger differences are likely to be observed for cases where the standard deviations become very large or where the data matrix is small. Such cases justify efforts to improve the efficiency of the modified algorithm.

5.4. Correlated errors

Figure 4(d) reveals some interesting characteristics of MLPCA when applied to cases of correlated errors. It is clear from the figure that the version of the MLPCA algorithm that incorporates error covariance provides the optimum model according to the maximum likelihood criterion and that PCA models are inferior in this regard. An additional analysis was also carried out using the standard MLPCA algorithm by assuming no correlation among the errors and using the diagonal elements of the full covariance matrix (Ω) for the variances. The models generated by this approach were not visibly any better than the PCA models, although the plot does not allow a direct comparison of these two sets of results. Further studies have shown that PCA and the standard MLPCA algorithm produce inferior results for correlated errors even when the standard deviations are the same for all measurements (in this case these two algorithms are equivalent). This indicates the importance of the error covariance information. Since many chemical measurements and data-preprocessing methods give rise to correlated errors, future studies need to be carried out to assess the importance of this contribution to model estimation and to improve the numerical reliability of the algorithm incorporating covariance.

5.5. Model quality

Although MLPCA generates models with smaller values of the objective function than does PCA, the key questions have yet to be answered. Are the MLPCA models closer to the true model and do they offer significant advantages over the PCA models? The second part of this question cannot be answered outside the context of particular applications, since the advantages gained by MLPCA will undoubtedly depend on the type and magnitude of errors involved as well as the intended use of the model (regression, mixture analysis, etc.). However, the first part of the question is readily answered using simulated data.

One way to compare the MLPCA and PCA models is to project the original vectors used to generate the error-free data onto the row space (U) or column space (V) of the model. The angle between the projected vector and the original vector then gives an indication of the agreement between the true

Table 4. Comparison of vector angle accuracies for PCA and MLPCA. Results are based on 100 replicates and uncertainties are given as standard deviations

Data set number	Rank	Mean angular deviation (PCA) (deg)		Mean angular deviation (MLPCA) (deg)	
		<i>U</i>	<i>V</i>	<i>U</i>	<i>V</i>
2	2	0.41±0.08	0.39±0.06	0.22±0.04	0.21±0.05
		0.42±0.08	0.41±0.06	0.22±0.04	0.23±0.04
4	3	3.1±0.6	3.4±0.6	2.3±0.4	3.0±0.5
		4.7±1.0	3.8±0.7	3.8±0.7	3.1±0.5
		3.0±0.7	3.5±0.6	2.3±0.4	2.9±0.4
7	2	0.10±0.05	0.16±0.05	0.023±0.012	0.10±0.04
		0.11±0.06	0.16±0.06	0.027±0.016	0.12±0.04

model and the fitted model. Mathematically, the angular deviations for the left-hand vectors are given by

$$\theta_i = \cos^{-1} \left(\frac{\mathbf{a}_i^T \mathbf{U} \mathbf{U}^T \mathbf{a}_i}{\|\mathbf{a}_i^T\| \cdot \|\mathbf{U} \mathbf{U}^T \mathbf{a}_i\|} \right) \quad (38)$$

where θ_i is the angular deviation of the left-hand vector \mathbf{a}_i from the space of the model. A similar expression can be used for the right-hand vectors. In order to be able to draw statistically valid conclusions, 100 replicates were run for data sets 2, 4 and 7. The results are reported in Table 4. Note that since equation (38) always gives positive values, the results in the table are the mean values of the absolute angular deviations. Also given are the standard deviations of the distributions.

Application of the *t*-test to the results in Table 4 clearly indicates that the MLPCA algorithm produces models with smaller angular deviations for all three data sets. The extent of improvement varies considerably with the nature of the data and the errors, however. Although these differences are statistically significant, the practical significance of the differences remain as a subject for future research.

To further assess the quality of MLPCA model estimation, the experimental data set, data set 6, was used. Of course, in this case, replicate data sets were unavailable, as were pure component spectra. However, the deviations of the concentration vectors from the model space could be measured. The concentrations for this data set were determined gravimetrically and so were accurately known. Table 5 shows that when the full wavelength range is used, the model obtained by MLPCA is far superior

Table 5. Angular deviations of concentration vectors from PCA space for data set 6

Wavelength range (nm)	Component	Angular deviation (deg)	
		PCA	MLPCA
400–2500	Toluene	13.7	0.127
	Chlorobenzene	13.6	0.135
	Heptane	25.8	0.688
700–1600	Toluene	0.281	0.281
	Chlorobenzene	0.359	0.359
	Heptane	0.792	0.792

to that obtained by PCA. This is not surprising, since PCA attempts to model the variance due to measurement errors in the high-absorbance regions. Normally in a situation like this one would preselect wavelengths prior to analysis by PCA. If the wavelength region used is 700–1600 nm, the variance is essentially uniform and PCA and MLPCA produce equivalent results. However, it will be noted that the angular deviation of the concentration vectors is smaller when MLPCA is used with the full wavelength region as opposed to the truncated data set, even though the error estimates in this case are only approximate. This illustrates that information can be lost when data are excluded from the modeling process. MLPCA uses the measurement error variance to optimize the amount of information extracted from the data and in that sense represents a significant advance in multivariate analysis.

6. CONCLUSIONS

In this work the theoretical foundations of MLPCA have been established using the framework of PCA (and SVD). By incorporating information about the measurement errors, the procedure has been shown to be optimal for principal component modeling in accordance with a maximum likelihood criterion. The algorithm presented here is particularly efficient in its use of alternating regression to achieve rapid convergence. Modifications to the algorithm also permit the incorporation of intercept terms consistent with a maximum likelihood model. Furthermore, generalization of the method allows the incorporation of correlated measurement errors. To our knowledge, this is the first time that a PCA procedure has been developed that has the capability of dealing with measurement error covariance. Results using simulated data show that the objective function minimized by the algorithm approximates a χ^2 -distribution with $(m-p)(n-p)$ degrees of freedom (in the absence of intercept terms) provided that the measurement error covariance matrix is known and the form of the model is correct. Results using simulated and experimental data also demonstrate that model estimation by MLPCA is superior to models produced by PCA in cases where non-uniform or correlated errors are present.

The practical implications of the theoretical aspects of MLPCA put forward here will no doubt be the subject of future research. This work has clearly demonstrated the positive features of MLPCA and answered many questions related to optimal scaling for PCA models. The advantages of MLPCA over PCA are balanced to some degree by the greater computational efficiency of the latter. As long as measurement errors are approximately uniform or are very small in the context of the intended application, PCA results may be sufficient, if suboptimal. However, there are many cases in chemistry where these conditions do not hold. Applications in a wide range of areas, including calibration, curve resolution and data fusion, can no doubt benefit from the optimal modeling features of MLPCA and are fertile grounds for future research. This work has demonstrated the importance of measurement error data for maximizing the information available from chemical data sets and MLPCA should serve as an important archetype for optimal modeling by PCA.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Center for Process Analytical Chemistry.

APPENDIX I: NOTATION AND LIST OF SYMBOLS

In general the conventions used in this paper are as follows. Matrices are represented as bold uppercase letters and column vectors are represented as bold lowercase letters. Italic upper- and lowercase letters are used for scalars. Greek and italic fonts are used with no particular pattern, but

where possible we have tried to adhere to symbols commonly used in the literature. Symbols which represent estimates of unknown quantities are designated with a caret. A matrix transpose is indicated by a superscript 'T' and the Euclidean norm of a vector by '||·||'. The Kronecker product of two matrices is indicated '⊗'.

A list of important symbols in the paper follows.

$\mathbf{1}_m$	$m \times 1$ vector of ones
a_{ij}	model coefficient (element of \mathbf{A})
\mathbf{a}_i	left-hand vector of true model
\mathbf{A}	(i) left-hand matrix of p -dimensional model (equation (2)) (ii) matrix of model coefficients (equation (7))
$\hat{\mathbf{A}}$	matrix of estimated model coefficients
\mathbf{B}	(i) right-hand matrix of p -dimensional model (equation (2)) (ii) matrix of model offsets (equation (30))
\mathbf{b}_j	column vector of \mathbf{B} (offsets)
\mathbf{c}	vector of column offsets for MLPCA model
\mathbf{d}	vector of row offsets for MLPCA model
\mathbf{E}	unfiltered measurement errors for \mathbf{X}
\mathbf{F}	matrix of filter coefficients ($mn \times mn$) for calculating $\mathbf{\Omega}$ of filtered data
\mathbf{G}_j	intermediate matrix in calculation of $dS^2/d\alpha_i$
\mathbf{H}_j	substitution equal to $(\mathbf{U}^T \mathbf{\Psi}_j^{-1} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{\Psi}_j^{-1}$
\mathbf{I}_n	$n \times n$ identity matrix
\mathbf{J}_i	derivative of rotation matrix \mathbf{T} with respect to angle α_i
\mathbf{K}	commutation matrix for \mathbf{X}
l	function minimized for maximum likelihood projection
L	probability density function for measurement vector \mathbf{x}
\mathbf{L}_i	exchange matrix with property $\mathbf{J}_i = \mathbf{L}_i \mathbf{T}_i$
m	number of rows in \mathbf{X}
n	number of columns in \mathbf{X}
p	rank of data matrix (pseudorank)
P	probability of observing a χ^2 -value less than a given value
\mathbf{P}_j	maximum likelihood projection matrix for \mathbf{x}_j
S^2	objective function
$\mathbf{T}_1, \mathbf{T}_2, \dots$	individual rotation matrices
\mathbf{T}	overall rotation matrix
u, s, v	elements of $\hat{\mathbf{U}}$, $\hat{\mathbf{S}}$ and $\hat{\mathbf{V}}$
$\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}$	estimated MLPCA matrices (SVD form)
$\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2$	upper and lower matrices of $\hat{\mathbf{U}}$ ($p \times p$ and $(m-p) \times p$)
$\hat{\mathbf{U}}_0, \hat{\mathbf{V}}_0$	initial estimates for $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$
\mathbf{U}, \mathbf{V}	supermatrices for $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$
$x_{ij}, \hat{x}_{ij}, x_{ij}^0$	elements of \mathbf{X} , $\hat{\mathbf{X}}$ and \mathbf{X}^0
$\mathbf{x}_j, \hat{\mathbf{x}}_j, \mathbf{x}_j^0$	column vectors of \mathbf{X} , $\hat{\mathbf{X}}$ and \mathbf{X}^0
$\hat{\mathbf{x}}_p, \mathbf{x}_p^0$	upper $p \times 1$ vectors of $\hat{\mathbf{x}}$ and \mathbf{x}^0
$\Delta \mathbf{x}_j$	residual vector for column j of \mathbf{X} ($= \mathbf{x}_j - \hat{\mathbf{x}}_j$)
\mathbf{X}	measurement data matrix
$\hat{\mathbf{X}}$	matrix of maximum likelihood estimates of \mathbf{X}^0
$\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2$	upper and lower matrices of $\hat{\mathbf{X}}$ ($p \times n$ and $(m-p) \times n$)
\mathbf{X}^0	matrix of true measurements
\mathbf{y}	vector of dependent variables in classical regression
α	significance level for hypothesis testing
α_i	rotation angle about axis i
β	vector of regression coefficients in classical regression
δ	vector of errors in \mathbf{y} in classical regression
\mathbf{e}_j	column vector of measurement errors for column j of \mathbf{X}
\mathbf{e}_{ij}	measurement error for x_{ij} after filtering
Φ_{ij}	matrix of filter coefficients for errors corresponding to measurement x_{ij}
γ	vector of row offsets for Mandel model

Γ^{inc}	incomplete gamma function
λ	convergence criterion for MLPCA algorithm
μ	grand mean for \mathbf{X}
ν	degrees of freedom
θ_i	angular deviation of eigenvector i from true model space
ρ	vector of column offsets for Mandel model
Σ_i	error covariance matrix for row i of \mathbf{X} (or column i of \mathbf{X}^T)
Ω	full error covariance matrix of $\text{vec}(\mathbf{X})$
Ξ	full covariance matrix of $\text{vec}(\mathbf{X}^T)$
Ψ_j	error covariance matrix for column vector j of \mathbf{X}
σ_{ij}^2	measurement variance for x_{ij}

APPENDIX II: DERIVATION OF MAXIMUM LIKELIHOOD PREDICTION EQUATION

Given a vector \mathbf{x} of length m of observed measurements and the corresponding error covariance matrix Ψ , the multivariate probability density function at \mathbf{x} is given by

$$L = \frac{1}{(2\pi)^{m/2} |\Psi|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mathbf{x}^0)^T \Psi^{-1} (\mathbf{x} - \mathbf{x}^0)] \tag{39}$$

where \mathbf{x}^0 represents the (unknown) vector of true values. The vector of maximum likelihood estimates of \mathbf{x}^0 for a given $\hat{\mathbf{A}}$, designated $\hat{\mathbf{x}}$, is obtained by maximizing the probability density function subject to the parametric model $\hat{\mathbf{x}} = \hat{\mathbf{A}}\hat{\mathbf{x}}_p$. This corresponds to minimizing the function

$$l = (\mathbf{x} - \hat{\mathbf{x}})^T \Psi^{-1} (\mathbf{x} - \hat{\mathbf{x}}) \tag{40}$$

with respect to $\hat{\mathbf{x}}$. Substitution of the parametric model gives

$$\begin{aligned} l &= (\mathbf{x} - \hat{\mathbf{A}}\hat{\mathbf{x}}_p)^T \Psi^{-1} (\mathbf{x} - \hat{\mathbf{A}}\hat{\mathbf{x}}_p) \\ &= \mathbf{x}^T \Psi^{-1} \mathbf{x} - \mathbf{x}^T \Psi^{-1} \hat{\mathbf{A}}\hat{\mathbf{x}}_p - \hat{\mathbf{x}}_p^T \hat{\mathbf{A}}^T \Psi^{-1} \mathbf{x} + \hat{\mathbf{x}}_p^T \hat{\mathbf{A}}^T \Psi^{-1} \hat{\mathbf{A}}\hat{\mathbf{x}}_p \end{aligned} \tag{41}$$

Using standard relations for derivatives of vectors,²² this gives

$$\frac{\partial l}{\partial \hat{\mathbf{x}}_p} = 0 - \hat{\mathbf{A}}^T \Psi^{-1} \mathbf{x} - \hat{\mathbf{A}}^T \Psi^{-1} \mathbf{x} + 2\hat{\mathbf{A}}^T \Psi^{-1} \hat{\mathbf{A}}\hat{\mathbf{x}}_p \tag{42}$$

Setting this equal to zero to find the minimum leads to

$$\hat{\mathbf{x}}_p = (\hat{\mathbf{A}}^T \Psi^{-1} \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T \Psi^{-1} \mathbf{x} \tag{43}$$

which is the same as equation (9) and leads directly to equation (10).

APPENDIX III: DERIVATIVES OF S^2

The calculation of the derivative of S^2 with respect to the rotation angles (in the absence of intercept terms) begins by finding the differential of S^2 .

$$\begin{aligned} dS^2 &= \sum_{j=1}^n d(\Delta \mathbf{x}_j^T \Psi_j^{-1} \Delta \mathbf{x}_j) \\ &= \sum [d(\Delta \mathbf{x}_j^T) \Psi_j^{-1} \Delta \mathbf{x}_j + \Delta \mathbf{x}_j^T \Psi_j^{-1} d(\Delta \mathbf{x}_j)] \end{aligned}$$

$$= 2 \sum \Delta \mathbf{x}_j^T \Psi_j^{-1} d(\Delta \mathbf{x}_j) \quad (44)$$

It will be assumed that the symmetric error covariance matrix Ψ is known for each \mathbf{x} . For convenience, from this point on, the subscript 'j' in equation (44) will be dropped but will be implied. Also, for simplicity we will use \mathbf{U} in place of $\hat{\mathbf{U}}$. We can write $\Delta \mathbf{x}$ as

$$\begin{aligned} \Delta \mathbf{x} &= [\mathbf{I} - \mathbf{U}(\mathbf{U}^T \Psi^{-1} \mathbf{U})^{-1} \mathbf{U}^T \Psi^{-1}] \mathbf{x} \\ &= [\mathbf{I} - \mathbf{TU}_0 [(\mathbf{TU}_0)^T \Psi^{-1} \mathbf{TU}_0]^{-1} (\mathbf{TU}_0)^T \Psi^{-1}] \mathbf{x} \\ &= [\mathbf{I} - \mathbf{TU}_0 (\mathbf{U}_0^T \mathbf{T}^T \Psi^{-1} \mathbf{TU}_0)^{-1} \mathbf{U}_0^T \mathbf{T}^T \Psi^{-1}] \mathbf{x} \end{aligned} \quad (45)$$

This gives

$$\begin{aligned} \Delta \mathbf{x}^T \Psi^{-1} d(\Delta \mathbf{x}) &= - \{ \Delta \mathbf{x}^T \Psi^{-1} (d\mathbf{T}) \mathbf{U}_0 (\mathbf{U}^T \Psi^{-1} \mathbf{U}^T)^{-1} \mathbf{U}^T \Psi^{-1} \mathbf{x} + \Delta \mathbf{x}^T \Psi^{-1} \mathbf{U} [d(\mathbf{U}_0^T \mathbf{T}^T \Psi^{-1} \mathbf{TU}_0)^{-1}] \mathbf{U}^T \Psi^{-1} \mathbf{x} \\ &\quad + \Delta \mathbf{x}^T \Psi^{-1} \mathbf{U} (\mathbf{U}^T \Psi^{-1} \mathbf{U}^T)^{-1} \mathbf{U}_0^T (d\mathbf{T}^T) \Psi^{-1} \mathbf{x} \} \end{aligned} \quad (46)$$

If we make the substitution $\mathbf{H} = (\mathbf{U}^T \Psi^{-1} \mathbf{U})^{-1} \mathbf{U}^T \Psi^{-1}$, equation (46) can be further expanded by conventional means¹⁸ to give

$$\begin{aligned} \Delta \mathbf{x}^T \Psi^{-1} d(\Delta \mathbf{x}) &= - [\Delta \mathbf{x}^T \Psi^{-1} (d\mathbf{T}) \mathbf{U}_0 \mathbf{H} \mathbf{x} - \mathbf{x}^T \Psi^{-1} \mathbf{U} \mathbf{H} (d\mathbf{T}) \mathbf{U}_0 \mathbf{H} \Delta \mathbf{x} \\ &\quad - \Delta \mathbf{x}^T \Psi^{-1} \mathbf{U} \mathbf{H} (d\mathbf{T}) \mathbf{U}_0 \mathbf{H} \mathbf{x} + \mathbf{x}^T \Psi^{-1} (d\mathbf{T}) \mathbf{U}_0 \mathbf{H} \Delta \mathbf{x}] \end{aligned} \quad (47)$$

The differential of the transformation matrix can be expanded as

$$d\mathbf{T} = (d\mathbf{T}_1) \mathbf{T}_2 \mathbf{T}_3 \dots \mathbf{T}_{m-1} + \mathbf{T}_1 (d\mathbf{T}_2) \mathbf{T}_3 \dots \mathbf{T}_{m-1} + \dots \quad (48)$$

where

$$d\mathbf{T}_i = \frac{d\mathbf{T}_i}{d\alpha_i} d\alpha_i = \mathbf{J}_i d\alpha_i \quad (49)$$

Here \mathbf{J}_i is the derivative of the rotation matrix corresponding to α_i . For example,

$$\mathbf{J}_1 = \begin{bmatrix} -\sin \alpha_1 & -\cos \alpha_1 & 0 & \dots & 0 \\ \cos \alpha_1 & -\sin \alpha_1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{J}_2 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & -\sin \alpha_2 & -\cos \alpha_2 & \dots & 0 \\ 0 & \cos \alpha_2 & -\sin \alpha_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad \text{etc.} \quad (50)$$

Furthermore, it is easily shown that

$$\mathbf{J}_i = \mathbf{L}_i \mathbf{T}_i \quad (51)$$

where

$$\mathbf{L}_1 = \begin{bmatrix} 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{L}_2 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad \text{etc.} \quad (52)$$

From this we can write

$$\begin{aligned}
 d\mathbf{T} &= \mathbf{J}_1(\mathbf{T}_2\mathbf{T}_3\dots) d\alpha_1 + \mathbf{T}_1\mathbf{J}_2(\mathbf{T}_3\mathbf{T}_4\dots) d\alpha_2 + \dots \\
 &= \mathbf{L}_1\mathbf{T}_1(\mathbf{T}_2\mathbf{T}_3\dots) d\alpha_1 + \mathbf{T}_1\mathbf{L}_2\mathbf{T}_2(\mathbf{T}_3\mathbf{T}_4\dots) d\alpha_2 + \dots \\
 &= \mathbf{L}_1\mathbf{T} d\alpha_1 + \mathbf{T}_1\mathbf{L}_2\mathbf{T}_1^{-1}\mathbf{T}_1\mathbf{T}_2(\mathbf{T}_3\mathbf{T}_4\dots) d\alpha_2 + \dots \\
 &= \mathbf{L}_1\mathbf{T} d\alpha_1 + (\mathbf{T}_1\mathbf{L}_2\mathbf{T}_1^T)\mathbf{T} d\alpha_2 + \dots \\
 &= \mathbf{G}_1\mathbf{T} d\alpha_1 + \mathbf{G}_2\mathbf{T} d\alpha_2 + \dots + \mathbf{G}_{m-1}\mathbf{T} d\alpha_{m-1}
 \end{aligned} \tag{53}$$

where

$$\mathbf{G}_i = (\mathbf{T}_1\mathbf{T}_2\dots\mathbf{T}_{i-1})\mathbf{L}_i(\mathbf{T}_{i-1}^T\mathbf{T}_{i-2}^T\dots\mathbf{T}_1) \tag{54}$$

Substitution of equation (53) into equation (47) and recognizing that $\mathbf{U} = \mathbf{T}\mathbf{U}_0$ and $\mathbf{P} = \mathbf{U}\mathbf{H}$ gives

$$\Delta\mathbf{x}^T\boldsymbol{\Psi}^{-1}d(\Delta\mathbf{x}) = - \sum (\Delta\mathbf{x}^T\boldsymbol{\Psi}^{-1}\mathbf{G}_i\mathbf{P}\mathbf{x} - \mathbf{x}^T\boldsymbol{\Psi}^{-1}\mathbf{P}\mathbf{G}_i\mathbf{P}\Delta\mathbf{x} - \Delta\mathbf{x}^T\boldsymbol{\Psi}^{-1}\mathbf{P}\mathbf{G}_i\mathbf{P}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Psi}^{-1}\mathbf{G}_i\mathbf{P}\Delta\mathbf{x}) d\alpha_i \tag{55}$$

This result leads to equation (20) in the paper. Although this equation is correct and faster than numerical evaluation of the derivatives, it is somewhat cumbersome. Each of the $m - 1$ parameters to be optimized requires the calculation and storage of an $m \times m$ matrix \mathbf{G}_i , which is the product of $2i - 1$ matrices that are $m \times m$. Even though these matrices are sparse, the calculations are still time-consuming and awkward. Some simplification of equation (55) is possible by examining the characteristics of \mathbf{G} . The matrix \mathbf{G}_i is antisymmetric with zero elements everywhere except for the first i elements of column $i + 1$ and row $i + 1$. If the rotation angles are small, the rotation matrices approach the identity matrix and a good approximation is

$$\mathbf{G}_i \approx \mathbf{L}_i \tag{56}$$

We have found that this approximation works well in practice, particularly if \mathbf{U}_0 is updated as convergence is approached so that the angles remain small.

REFERENCES

1. R. J. Larsen and M. L. Marx, *An Introducton to Mathematical Statistics and Its Applications*, 2nd edn, Prentice-Hall, Englewood Cliffs, NJ (1986).
2. P. Paatero and U. Tapper, *Chemometrics Intell. Lab. Syst.* **18**, 183 (1993).
3. R. N. Cochran and F. H. Horne, *Anal. Chem.* **49**, 846 (1977).
4. H. R. Halvorson, *Biophys. Chem.* **14**, 177 (1984).
5. V. Simeon and D. Pavkovic, *J. Chemometrics* **6**, 257 (1992).
6. K. R. Gabriel and S. Zamir, *Technometrics* **21**, 489 (1979).
7. P. Paatero and U. Tapper, *Environmetrics* **5**, 111 (1994).
8. D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method*, 2nd edn, p. 109, Butterworths, London (1971).
9. K. G. Jöreskog, in *Systems Under Indirect Observation, Part I*, ed. by K. G. Jöreskog and H. Wold, Chap. 4, North-Holland, Amsterdam (1982).
10. K. A. Bollen, *Structural Equations with Latent Variables*, Wiley, New York (1989).
11. P. Moens, P. De Volder, R. Hoogewijs, F. Callens and R. Verbeeck, *J. Magn. Reson. A*, **101**, 1 (1993).
12. P. Persoone, R. De Gryse and P. De Volder, *J. Electron Spectrosc. Relat. Phenom.* **71**, 225 (1995).
13. S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem*, SIAM, Philadelphia, PA (1991).
14. W. A. Fuller, *Measurement Error Models*, Wiley, New York (1987).
15. M. S. Bartlett, *Br. J. Psychol.* **28**, 97 (1937).
16. M. S. Bartlett, *Nature* **141**, 609 (1938).

17. L. J. Gleaser and H. Yang, *Anal. Chim. Acta* **277**, 405 (1993).
18. J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, Chichester (1988).
19. J. Mandel, *Technometrics* **13**, 1 (1971).
20. T. Dean, B. Kowalski and R. Pell, *Appl. Spectrosc.* submitted.
21. W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in FORTRAN*, 2nd edn, p. 215, Cambridge University Press, New York (1992).
22. W. H. Beyer (ed.), *CRC Handbook of Tables for Probability and Statistics*, p. 26, Chemical Rubber Co., Cleveland, OH (1996).