

Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations

Peter D. Wentzell^{*}, Mitchell T. Lohnes

Trace Analysis Research Centre, Department of Chemistry, Dalhousie University, Halifax, NS, Canada B3H 4J3

Received 13 July 1997; revised 1 December 1997; accepted 12 January 1998

Abstract

Procedures to compensate for correlated measurement errors in multivariate data analysis are described. These procedures are based on the method of maximum likelihood principal component analysis (MLPCA), previously described in the literature. MLPCA is a decomposition method similar to conventional PCA, but it takes into account measurement uncertainty in the decomposition process, placing less emphasis on measurements with large variance. Although the original MLPCA algorithm can accommodate correlated measurement errors, two drawbacks have limited its practical utility in these cases: (1) an inability to handle rank deficient error covariance matrices, and (2) demanding memory and computational requirements. This paper describes two simplifications to the original algorithm that apply when errors are correlated only within the rows of a data matrix and when all of these row covariance matrices are equal. Simulated and experimental data for three-component mixtures are used to test the new methods. It was found that inclusion of error covariance information via MLPCA always gave results which were at least as good and normally better than PCA when the true error covariance matrix was available. However, when the error covariance matrix is estimated from replicates, the relative performance depends on the quality of the estimate and the degree of correlation. For experimental data consisting of mixtures of cobalt, chromium and nickel ions, maximum likelihood principal components regression showed an improvement of up to 50% in the cross-validation error when error covariance information was included. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Principal component analysis; Multivariate calibration; Errors; Maximum likelihood; Covariance

Contents

1. Introduction	66
2. Theory	67
2.1. MLPCA	67
2.2. Correlated measurement errors	69

^{*} Corresponding author. Tel.: +1-9024943305; Fax: +1-9024941310; E-mail: wentzell@chem1.chem.dal.ca

2.3. Simplification 1—row correlations only	70
2.4. Simplification 2—identical row correlations only	71
3. Experimental	73
3.1. Data sets	73
3.2. Computational aspects	74
4. Results and discussion	75
4.1. Simulated data	75
4.2. Experimental data	80
5. Conclusions	83
Acknowledgements.	83
Appendix A.	83
Appendix B.	84
References	85

1. Introduction

In chemometrics, principal component analysis (PCA) is perhaps the most widely used technique for the treatment of multivariate chemical data. It is especially valuable for problems of dimensionality reduction, rank estimation, modeling, and mixture analysis. Many methods have evolved with PCA at their core (e.g., principal components regression, SIMCA) and, although several variations of the basic algorithm exist (e.g., implementation by singular value decomposition and kernel algorithms), it is a testament to the utility of the method that it has withstood the test of time and remains fundamentally unchanged. However, a persistent weakness of PCA in chemical applications has been its inability to incorporate knowledge of measurement uncertainty into the analysis of experimental data. The utility of PCA derives from its ability to model variance in the data, but unfortunately it is unable to distinguish between the variance associated with the measurement process and systematic variance due to chemical factors. This is especially a problem when the measurement uncertainties have non-uniform variances and, over the years, a variety of ad hoc scaling and variable se-

lection methods have been developed in an attempt to address this issue. These approaches are often less than satisfactory, however, and until recently no general method was available to handle estimates of measurement uncertainty in an optimal fashion.

A breakthrough in this area came in 1994 when Paatero and Tapper [1] first described a technique they refer to as ‘positive matrix factorization’ (PMF). Although the results of this method are not directly interchangeable with PCA, the two techniques are similar in the sense that both provide a low dimensional model for points in a higher dimensional space. PMF has a number of useful features, but the one of interest here is its ability to account for estimates of measurement uncertainty in an optimal way. Paatero and Tapper were not the first (or last) to use measurement uncertainties in this manner (see for example Refs. [2,3]), but they were responsible for resurrecting the idea in the chemometrics/environmental literature. Unfortunately, their approach has not yet been widely adopted, probably because the technique is not readily associated with PCA and the algorithm used in the original work was not clearly defined. More recently, our group has independently developed a related technique which is referred to as

‘maximum likelihood principal component analysis’ (MLPCA) because the incorporation of measurement variance information in the model estimation is optimal in a maximum likelihood sense [4]. This approach has a number of advantages over PMF. First, the results are directly identifiable with the results from conventional PCA, which makes its adaptation into existing methods relatively straightforward. For example, the technique has recently been used to improve traditional multivariate calibration methods [5] and deal with the problem of incomplete data sets [6]. A second advantage of MLPCA is that the basic algorithm is an alternating least squares procedure which is very compact and has excellent convergence characteristics. The third, and perhaps most important feature of MLPCA, is that, unlike PMF and other similar approaches, it provides a method to accommodate correlated measurement errors.

The presence of covariance among measurement errors in chemistry is a ubiquitous phenomenon, with sources ranging from the temporal correlations of pump noise in chromatography to the spatial correlations of array detectors in spectroscopy. Furthermore, many kinds of signal processing methods, particularly electronic or digital smoothing or derivative filters, can give rise to correlations in measurement noise. No characterization of noise is complete without a description of these correlations, which are often expressed in the form of a noise power spectrum (Fourier transform). For example, the term ‘flicker noise’ is often used to describe signals with low-frequency correlations in the time domain. Perhaps a more general description of measurement error correlations, and the one that will be used in this work, is the one provided by the measurement error covariance matrix. This matrix describes the relationships among all of the errors in a series of measurements. In practice, the error covariance matrix, like the noise power spectrum, is seldom known exactly, but approximations to it can be very useful.

Despite the fact that measurement error correlations are almost universal for multivariate measurements, there have been very few, if any, chemometric techniques that have taken this into account. Historically, scalar quantities obtained from ‘zero-order’ instruments normally result in independent measurement errors from sample to sample, and this assumption of independent errors has continued to higher-

order instruments where it is no longer valid. The development of MLPCA now provides a framework from which the importance of correlated measurement errors can be assessed.

It is important to recognize that MLPCA is not just a variation of PCA but is a more general theory of multivariate modeling. PCA is a special case of this more general view that applies when measurement error variances are uniform and uncorrelated. The ability of MLPCA to handle correlated measurement errors is a unique characteristic of this method and demonstrates the generality of the underlying principles. Although the original paper on MLPCA [4] outlined the theory for dealing with correlated errors and demonstrated its validity through simulations on small matrices, there are two real limitations to the practical application of these principles. First, the practical estimation of error covariance matrices typically involves a finite number of replicates, which results in a rank deficient matrix unless the number of replicates is greater than the number of columns. Because this matrix is inverted in the MLPCA procedure, problems arise. The second problem is one of practical limitations on memory. In the most general case, the full covariance matrix for a matrix of N elements will contain N^2 elements, and this rapidly exceeds the storage capacity of most machines for even moderate arrays. Even if the full covariance matrix is sparse, storage requirements can still be large and it would be useful to find methods to reduce it further.

In this paper, the theoretical and practical aspects of MLPCA for correlated measurement errors are examined in more detail. Methods for improving numerical stability and reducing the storage requirements are examined. These techniques are tested on experimental and simulated data sets to demonstrate their utility and reliability.

2. Theory

2.1. MLPCA

Although a thorough theoretical treatment of MLPCA has been given elsewhere [4], a brief description will be presented here for completeness. We

begin by considering a $m \times n$ matrix, \mathbf{X} , of experimental measurements. For example, \mathbf{X} could represent spectra measured for m samples at n wavelengths. We will assume that the matrix consists of the ‘true’ (and unknown) measurements (\mathbf{X}^0) corrupted by measurement errors, \mathbf{E} , i.e.:

$$\mathbf{X} = \mathbf{X}^0 + \mathbf{E}. \quad (1)$$

Furthermore, we assume that \mathbf{X}^0 has a rank of p as a result of the fact that it is a linear combination of p underlying factors, i.e.:

$$\mathbf{X}^0 = \mathbf{A}\mathbf{B} \quad (2)$$

where \mathbf{A} is $m \times p$ and \mathbf{B} is $p \times n$. For example, \mathbf{A} may represent concentrations of the p components in the m samples and \mathbf{B} may represent the corresponding molar absorptivities at n wavelengths. Generally, \mathbf{A} , \mathbf{B} and p are unknown, but PCA can be applied in an effort to determine the rank and the subspace containing \mathbf{A} and \mathbf{B} . Typically this is done by singular value decomposition (SVD) of the experimental data which gives the following:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad (3)$$

Here, if \mathbf{X} is $m \times n$ with $m \leq n$, then \mathbf{U} is $m \times m$, \mathbf{S} is $m \times m$, and \mathbf{V} is $n \times m$. The columns of \mathbf{U} and \mathbf{V} are orthonormal, and \mathbf{S} is the diagonal matrix of singular values. Unlike \mathbf{X}^0 , \mathbf{X} contains random measurement errors and is typically full rank in a numerical sense. However, the chemical rank, or pseudorank, of \mathbf{X} is only p . Normally, \mathbf{X}^0 is estimated by reconstructing the data from a truncated SVD in which only the first p principal components are retained to give:

$$\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T \quad (4)$$

where $\hat{\mathbf{U}}$ is $m \times p$, $\hat{\mathbf{S}}$ $p \times p$ and $\hat{\mathbf{V}}$ is $n \times p$. Typically, the reconstruction using p principal components is compared to the original data to see if it is sufficient to describe the measurements within experimental error. In this way, the pseudorank of \mathbf{X} can be estimated.

In the above procedure, there was no consideration of the characteristics of the error matrix, \mathbf{E} , in the calculation. In contrast to conventional PCA, MLPCA incorporates such information into the decomposition process. In one of the simplest scenarios,

we will assume that the measurement errors are drawn from independent normal distributions with zero means and known variances; that is:

$$e_{ij} \sim N(0, \sigma_{ij}^2) \quad (5)$$

where e_{ij} is an element of \mathbf{E} . (The philosophy of maximum likelihood estimation is not restricted to normally distributed errors, but such an assumption does simplify the mathematics.) Exact variances, of course, are not generally known and must be estimated from a finite number of samples, but it has been shown elsewhere [5] that MLPCA can still be used effectively with the variance estimates.

In the case of independent errors, the measurement error variance can be described in terms of a series of n diagonal $m \times m$ column covariance matrices, Ψ_j :

$$\Psi_j = \begin{bmatrix} \sigma_{1j}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2j}^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{mj}^2 \end{bmatrix} \quad (6)$$

Alternatively, the same information is contained in the m diagonal $n \times n$ row covariance matrices, Σ_i :

$$\Sigma_i = \begin{bmatrix} \sigma_{i1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{i2}^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{in}^2 \end{bmatrix} \quad (7)$$

In MLPCA, the objective function to be minimized (S^2) is a summation of weighted squared residuals given by:

$$\begin{aligned} S^2 &= \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ij} - \hat{x}_{ij})^2}{\sigma_{ij}^2} \\ &= \sum_{j=1}^n (\mathbf{x}_{\cdot j} - \hat{\mathbf{x}}_{\cdot j})^T \Psi_j^{-1} (\mathbf{x}_{\cdot j} - \hat{\mathbf{x}}_{\cdot j}) \\ &= \sum_{i=1}^m (\mathbf{x}_{i \cdot} - \hat{\mathbf{x}}_{i \cdot}) \Sigma_i^{-1} (\mathbf{x}_{i \cdot} - \hat{\mathbf{x}}_{i \cdot})^T \end{aligned} \quad (8)$$

Here $\mathbf{x}_{\cdot j}$ represents a column vector of \mathbf{X} and $\mathbf{x}_{i \cdot}$ represents a row vector. Likewise, the $\hat{\mathbf{x}}$ s indicate es-

timates of ‘true’ measurements derived from the model parameters.

Although Eq. (8) describes the minimization criterion, it does not specify what parameters are to be minimized. Considering Eq. (4), the columns of \hat{U} represent a p -dimensional orthogonal basis for the n -points in the m -dimensional row space. Alternatively, we can view the columns of \hat{V} as a p -dimensional basis for the m points in the n -dimensional column space. In either case, we seek the optimum rotation of the p vectors describing the subspace that will minimize S^2 in Eq. (8). In evaluating Eq. (8) for a trial model (\hat{U} or \hat{V}) we need \hat{x}_{ij} , the projections of x_{ij} into the subspace of the model. In PCA, an orthogonal projection is used. In MLPCA, a maximum likelihood projection is used. Depending on whether one is working in the row space or column space, Eq. (9) or Eq. (10) is used for the ML projection:

$$\hat{x}_{.j} = \hat{U}(\hat{U}^T \Psi_j^{-1} \hat{U})^{-1} \hat{U}^T \Psi_j^{-1} x_{.j} \quad (9)$$

$$\hat{x}_i = x_i \Sigma_i^{-1} \hat{V}(\hat{V} \Sigma_i^{-1} \hat{V})^{-1} \hat{V}^T \quad (10)$$

Note that \hat{U} and \hat{V} in these equations denote trial solutions in general and are not necessarily the results of conventional PCA. These projections are weighted by the measurement uncertainties in such a way that, for a given trial model, estimates of the true measurements will be optimal in a maximum likelihood sense. It is easily shown that, when all of the measurements have uniform variance, Eqs. (9) and (10) lead to the usual orthogonal projection employed in conventional PCA, i.e. $\hat{x}_{.j} = \hat{U} \hat{U}^T x_{.j}$ and $\hat{x}_i = x_i \hat{V} \hat{V}^T$. This is consistent with the widely known fact that PCA minimizes the sum of the squares of the orthogonal residuals, and conventional PCA is a maximum likelihood modeling method under equal measurement variance conditions. However, when these conditions are violated, this is no longer true.

In principle, Eq. (8) can be minimized by finding the optimum rotation of either \hat{U} or \hat{V} . The remaining matrices (\hat{S} and \hat{V} or \hat{U} and \hat{S}) can then be easily found by applying conventional SVD to the ML estimates in \hat{X} (since these are restricted to a p -dimensional subspace). In practice, however, this approach involves a multiparameter nonlinear optimization which is often painfully slow and prone to local minima. The solution to this problem was described in an

earlier work [4] and utilizes the elegant simplicity of an alternating least squares (ALS) algorithm. Briefly, this solution involves alternating between the row and column spaces. The projections for a trial solution in one space are used to estimate the solution in the alternative space, and this process is repeated until convergence is achieved. In our experience, this algorithm has proven to be relatively fast (although still orders of magnitude slower than standard SVD) and remarkably reliable.

2.2. Correlated measurement errors

The inclusion of correlated errors into MLPCA is problematic in three ways: (1) theoretically, in that non-diagonal covariance matrices must now be considered; (2) algorithmically, in that swapping between row and column spaces is no longer trivial; and (3) numerically, in that matrices are larger and, in practical applications, may be numerically unstable. Each of these problems will be addressed in turn.

In real applications of multivariate analysis, correlations in measurement errors are likely to exist more often than not, although this fact is generally ignored because (a) the covariance information is generally unavailable, and (b) even if it were available, there are few analysis methods capable of dealing with it. These two problems reinforce each other, however, and it is important to develop general theories for dealing with correlated errors to break the cycle.

It is possible to distinguish several cases of correlated measurement errors. In some instances, correlations may only exist among measurements within each row or each column of the data matrix. This may occur, for example, on an absorption spectrometer where measurement errors on adjacent channels are correlated by source flicker noise or cell positioning errors, but there are no error correlations from sample to sample. In another scenario, such as a fluorescence excitation–emission spectrum, we might expect measurement error correlations to exist over the entire matrix and not be restricted to rows or columns. It is this general case that will be considered first and then simplified to other scenarios.

In the general case, it is obvious that the simple row and column covariance matrices, Σ and Ψ , will be insufficient to describe the relationships among all of the measurement errors. These matrices describe

only $mn(m+n)/2$ relationships (the factor of 2 is the result of the symmetry), while a full description of covariance involves $mn(mn+1)/2$ interactions. Therefore, we need a new way to describe covariance and this is best accomplished by introducing the *vec* notation. The full covariance matrix, $\mathbf{\Omega}$, is defined by:

$$\mathbf{\Omega} = E \left[\text{vec}(\mathbf{X} - \mathbf{X}^0) \cdot (\text{vec}(\mathbf{X} - \mathbf{X}^0))^T \right] \quad (11)$$

Here E denotes an expectation value. The *vec* operator converts the $m \times n$ matrix $(\mathbf{X} - \mathbf{X}^0)$ into an $mn \times 1$ vector by concatenating the columns in sequence [7]. Thus, $\mathbf{\Omega}$ is an $mn \times mn$ symmetric matrix. With this definition, the ML projection and objective function have been shown to be given by Eqs. (12) and (13), respectively:

$$\text{vec}(\hat{\mathbf{X}}) = \hat{\mathcal{Z}} (\hat{\mathcal{Z}}^T \mathbf{\Omega}^{-1} \hat{\mathcal{Z}})^{-1} \hat{\mathcal{Z}}^T \mathbf{\Omega}^{-1} \text{vec}(\mathbf{X}) \quad (12)$$

$$S^2 = (\text{vec}(\mathbf{X} - \hat{\mathbf{X}}))^T \mathbf{\Omega}^{-1} \text{vec}(\mathbf{X} - \hat{\mathbf{X}}) \quad (13)$$

where:

$$\hat{\mathcal{Z}} = \mathbf{I}_n \otimes \hat{\mathbf{U}} \quad (14)$$

Eq. (14) uses the Kronecker product (\otimes) which generates an $mn \times np$ block diagonal matrix, with $\hat{\mathbf{U}}$ replicated along the diagonal.

As before, S^2 can be minimized by rotation of the p vectors in $\hat{\mathbf{U}}$ in the m -dimensional row space of the original matrix. While this solves the theoretical problem of incorporating correlated errors into MLPCA, it raises an algorithmic one related to how to implement the ALS solution. This problem is easily addressed by vectorizing \mathbf{X}^T . The important equations are:

$$\mathbf{\Xi} = E \left[\text{vec}\{(\mathbf{X} - \mathbf{X}^0)^T\} \cdot \left[\text{vec}\{(\mathbf{X} - \mathbf{X}^0)^T\} \right]^T \right] \quad (15)$$

$$\text{vec}(\hat{\mathbf{X}}^T) = \hat{\mathcal{Z}} (\hat{\mathcal{Z}}^T \mathbf{\Xi}^{-1} \hat{\mathcal{Z}})^{-1} \hat{\mathcal{Z}}^T \mathbf{\Xi}^{-1} \text{vec}(\mathbf{X}^T) \quad (16)$$

$$S^2 = \left[\text{vec}\{(\mathbf{X} - \hat{\mathbf{X}})^T\} \right]^T \mathbf{\Xi}^{-1} \text{vec}\{(\mathbf{X} - \hat{\mathbf{X}})^T\} \quad (17)$$

$$\hat{\mathcal{Z}} = \mathbf{I}_m \otimes \hat{\mathbf{V}} \quad (18)$$

In these equations, $\mathbf{\Xi}$ is the full covariance matrix for \mathbf{X}^T and contains the same information as $\mathbf{\Omega}$ in a different arrangement. The relationship between $\mathbf{\Omega}$ and $\mathbf{\Xi}$ is given by the commutation matrix, \mathbf{K} , which is an $mn \times mn$ permutation matrix with mn non-zero elements [7]. In this case, \mathbf{K} is defined by:

$$\mathbf{K}_{a,b_i} = 1 \quad (19)$$

where:

$$\mathbf{a} = [1 \ 2 \ 3 \ \dots \ mn]^T = \text{vec}(\mathbf{A}) \quad (20)$$

$$\mathbf{b} = \text{vec}(\mathbf{A}^T) \quad (21)$$

and \mathbf{A} is an $m \times n$ matrix such that $\mathbf{A}_{ij} = i + m(j-1)$. The remaining elements of \mathbf{K} are zero. With this definition we have the following:

$$\mathbf{\Xi} = \mathbf{K} \mathbf{\Omega} \mathbf{K}^T \quad (22)$$

The general algorithm for implementing MLPCA with correlated measurement errors is given in Ref. [4] and has been tested using simulated data. Although the general case is important from a theoretical standpoint, it is of limited practical significance because of the size of the matrices involved. For example, a 100×100 emission-excitation matrix would require storage of about 5×10^7 elements for the full covariance matrix (accounting for symmetry). The problems of inverting a $10^4 \times 10^4$ matrix would also have to be addressed. Fortunately, real chemical problems often involve simplifications that can make the numerical aspects more tractable. Such simplifications are discussed in Sections 2.3 and 2.4.

2.3. Simplification 1—row correlations only

In many chemical applications, it may be reasonable to assume that measurement errors are correlated only along the rows or only along the columns. Traditionally, in calibration problems, individual spectra form the rows of \mathbf{X} . In the case of natural calibration or a well-designed experiment, there should be no correlations in the measurement errors for different samples, so we will treat the situation where correlations exist only along the rows. Obviously, the arguments can be easily modified to address the case of column covariance only, or the matrix can be transposed, so that situation will not be handled separately.

In the case of error covariance in the rows only, all of the covariance information is contained in the $n \times n$ row covariance matrices, Σ , defined by the following:

$$\Sigma_i = E[(\mathbf{x}_{i.} - \mathbf{x}_{i.}^0)^T \cdot (\mathbf{x}_{i.} - \mathbf{x}_{i.}^0)] \quad (23)$$

where $\mathbf{x}_{i.}$ and $\mathbf{x}_{i.}^0$ are row vectors of \mathbf{X} and \mathbf{X}^0 . The full covariance matrix, Ξ , will now be block diagonal, consisting of m diagonal units of dimensions $n \times n$. This reduces the number of non-zero elements from $m^2 n^2$ to mn^2 . While this is still a large number in most practical applications, it makes a critical difference in many cases, permitting Ξ to be stored as a sparse matrix. Furthermore, the block diagonal form allows Ξ to be inverted by inversion of the individual covariance blocks, i.e.:

$$\Xi^{-1} = \begin{bmatrix} \Sigma_1^{-1} & & & \\ & \Sigma_2^{-1} & & \\ & & \ddots & \\ & & & \Sigma_m^{-1} \end{bmatrix} \quad (24)$$

This improves the numerical stability of the algorithm. The inverse of the companion covariance matrix can be obtained by using the commutation matrix discussed in Section 2.2:

$$\Omega^{-1} = \mathbf{K}^T \Xi^{-1} \mathbf{K} \quad (25)$$

Although the above simplifications make the implementation of the covariance algorithm more feasible, there is still one more practical consideration which arises from the initial estimation of the covariance matrix. Although in some cases, there may be a theoretical basis for estimating the error covariance matrix, most real applications will require that it is estimated from a set of replicates by using the following equation:

$$\hat{\Sigma}_i = \frac{1}{(q-1)} \sum_{k=1}^q ((\mathbf{x}_{i.})_k - \bar{\mathbf{x}}_{i.})^T \cdot ((\mathbf{x}_{i.})_k - \bar{\mathbf{x}}_{i.}) \quad (26)$$

Here $(\mathbf{x}_{i.})_k$ indicates the k th replicate and $\bar{\mathbf{x}}_{i.}$ is the average of the q replicates (as before, both are row vectors). Unfortunately, this equation will lead to an $n \times n$ matrix with a maximum rank of $q - 1$. Therefore, if the number of replicates is smaller than the

number of channels, as is usually the case, $\hat{\Sigma}_i$ will be rank deficient and cannot be inverted. Often, use of the pseudo-inverse can remedy problems of singular matrices, but this will not work in this case. To illustrate, consider the simple case of regression in two dimensions where there are no errors in \mathbf{X} , i.e.:

$$\Sigma = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} \quad (27)$$

The projection of a point for a straight line model with a slope of 2 ($\mathbf{V}^T = [1/\sqrt{5} \ 2/\sqrt{5}]$) should be:

$$[\hat{x} \ \hat{y}] = [x \ y] \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix} = [x \ 2x] \quad (28)$$

The pseudo-inverse of Σ in this case is given by the following:

$$\Sigma^+ = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_y^{-2} \end{bmatrix} \quad (29)$$

which, by Eq. (10), gives a projection of:

$$[\hat{x} \ \hat{y}] = [x \ y] \begin{bmatrix} 0 & 0 \\ 0.5 & 1 \end{bmatrix} = [0.5y \ y] \quad (30)$$

This is a horizontal projection rather than a vertical projection, which is obviously incorrect. To solve this problem, we can instead add a small value to the diagonal elements of Σ prior to its inversion, similar to the approach used in ridge regression [8]. This value should be large enough to ensure numerical stability but small enough to avoid significantly perturbing the covariance matrix. We have found that the following equation works well:

$$\hat{\Sigma}'_i = \hat{\Sigma}_i + \mathbf{I}_n \|\hat{\Sigma}_i\| \cdot \varepsilon \cdot n \cdot 100 \quad (31)$$

In this equation, ε represents the machine precision.

2.4. Simplification 2—identical row correlations only

The procedure described above is effective for many problems, but there are ways to improve it further. In certain cases where the row covariance assumption applies, it may be possible to make the additional assumption that the error covariance matrix is the same for each row. This is reasonable, for example, in cases where the spectral measurement errors are independent of the magnitude of the signal.

An additional advantage is that estimates of the covariance matrix obtained from replicates of each sample can be pooled to obtain a better overall estimate that is given by the following:

$$\hat{\Sigma}_{\text{pooled}} = \frac{1}{m} \sum_{i=1}^m \hat{\Sigma}_i \quad (32)$$

The improvement in the estimation of Σ in this manner may even offset errors arising from violations in the assumption of identical covariance matrices.

Based on this assumption, a number of practical advantages in storage and speed can be gained. The block diagonal form of matrices in the row space make it easy to demonstrate that the ML projection in this space is:

$$\hat{\mathbf{X}} = \mathbf{X} \Sigma^{-1} \hat{\mathbf{V}} (\hat{\mathbf{V}}^T \Sigma^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \quad (33)$$

This is analogous to Eq. (10), except that it is now possible to project the entire matrix at once rather than one row vector at a time since Σ does not

Table 1

MLPCA algorithm for identical row error covariance matrices

(1) Given an $m \times n$ data matrix \mathbf{X} , a corresponding $n \times n$ row error covariance matrix Σ , and a model rank p , first check to see if Σ is rank deficient ($\text{rank}(\Sigma) < n$). If so adjust Σ according to:

$$\Sigma \leftarrow \Sigma + \mathbf{I}_n \|\Sigma\| \cdot \varepsilon \cdot n \cdot 100 \quad (\text{T-1})$$

where ε is the machine precision.

(2) Decompose \mathbf{X} by SVD and truncate the solution to rank p (designated as $\text{svd}(\mathbf{X}, p)$).

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\mathbf{X}, p) \quad (\text{T-2})$$

(3) Obtain maximum likelihood estimates of \mathbf{X} using $\hat{\mathbf{V}}$ and calculate the objective function, S_1^2 .

$$\hat{\mathbf{X}} = \mathbf{X} \Sigma^{-1} \hat{\mathbf{V}} (\hat{\mathbf{V}}^T \Sigma^{-1} \hat{\mathbf{V}})^{-1} \hat{\mathbf{V}}^T \quad (\text{T-3})$$

$$S_1^2 = \text{tr}[(\mathbf{X} - \hat{\mathbf{X}})^T \Sigma^{-1} (\mathbf{X} - \hat{\mathbf{X}})] \quad (\text{T-4})$$

(4) Decompose $\hat{\mathbf{X}}$ by SVD and obtain maximum likelihood estimates of \mathbf{X} using $\hat{\mathbf{U}}$. Calculate the objective function S_2^2 .

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\hat{\mathbf{X}}, p) \quad (\text{T-5})$$

$$\hat{\mathbf{X}} = \hat{\mathbf{U}} \hat{\mathbf{U}}^T \mathbf{X} \quad (\text{T-6})$$

$$S_2^2 = \text{tr}[(\mathbf{X} - \hat{\mathbf{X}})^T \Sigma^{-1} (\mathbf{X} - \hat{\mathbf{X}})] \quad (\text{T-7})$$

(5) Decompose $\hat{\mathbf{X}}$ by SVD.

$$[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}] \leftarrow \text{svd}(\hat{\mathbf{X}}, p) \quad (\text{T-8})$$

(6) Calculate the convergence parameter, λ .

$$\lambda = |(S_1^2 - S_2^2) / S_2^2| \quad (\text{T-9})$$

If λ is below the convergence limit, end. Otherwise, return to step 3.

change. The objective function is then calculated from the following:

$$S^2 = \sum_{i=1}^m (\mathbf{x}_i - \hat{\mathbf{x}}_i) \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \quad (34)$$

As usual, if we wish to use the ALS algorithm, we must find an analog to Eq. (33) in the column space. This can be done if we first recognize that:

$$\boldsymbol{\Xi}^{-1} = \mathbf{I}_m \otimes \boldsymbol{\Sigma}^{-1} \quad (35)$$

Using the properties of the commutation matrix (Ref. [7], p. 47), we can also write the following:

$$\begin{aligned} \boldsymbol{\Omega}^{-1} &= \mathbf{K}^T \boldsymbol{\Xi}^{-1} \mathbf{K} \\ &= \mathbf{K}^T (\mathbf{I}_m \otimes \boldsymbol{\Sigma}^{-1}) \mathbf{K} \\ &= \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_m \end{aligned} \quad (36)$$

Substituting into Eq. (12) and using the properties of the Kronecker product [9] gives:

$$\begin{aligned} \text{vec}(\hat{\mathbf{X}}) &= (\mathbf{I}_n \otimes \hat{\mathbf{U}}) \left[(\mathbf{I}_n \otimes \hat{\mathbf{U}})^T (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_m) (\mathbf{I}_n \otimes \hat{\mathbf{U}}) \right]^{-1} \\ &\quad \times (\mathbf{I}_n \otimes \hat{\mathbf{U}})^T (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_m) \text{vec}(\mathbf{X}) \\ &= (\mathbf{I}_n \otimes \hat{\mathbf{U}}) \left[(\mathbf{I}_n \otimes \hat{\mathbf{U}}^T) (\boldsymbol{\Sigma}^{-1} \otimes \hat{\mathbf{U}}) \right]^{-1} \\ &\quad \times (\boldsymbol{\Sigma}^{-1} \otimes \hat{\mathbf{U}}^T) \text{vec}(\mathbf{X}) \\ &= (\mathbf{I}_n \otimes \hat{\mathbf{U}}) \left[\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_p \right]^{-1} (\boldsymbol{\Sigma}^{-1} \otimes \hat{\mathbf{U}}^T) \text{vec}(\mathbf{X}) \\ &= (\mathbf{I}_n \otimes \hat{\mathbf{U}}) (\boldsymbol{\Sigma} \otimes \mathbf{I}_p) (\boldsymbol{\Sigma}^{-1} \otimes \hat{\mathbf{U}}^T) \text{vec}(\mathbf{X}) \\ &= (\boldsymbol{\Sigma} \otimes \hat{\mathbf{U}}) (\boldsymbol{\Sigma}^{-1} \otimes \hat{\mathbf{U}}^T) \text{vec}(\mathbf{X}) \\ &= (\mathbf{I}_n \otimes \hat{\mathbf{U}} \hat{\mathbf{U}}^T) \text{vec}(\mathbf{X}) \\ &= \text{vec}(\hat{\mathbf{U}} \hat{\mathbf{U}}^T \mathbf{X}) \\ &= \text{vec}(\hat{\mathbf{U}} \hat{\mathbf{U}}^T \mathbf{X}) \end{aligned} \quad (37)$$

or:

$$\hat{\mathbf{X}} = \hat{\mathbf{U}} \hat{\mathbf{U}}^T \mathbf{X} \quad (38)$$

The result is that, when row covariance matrices are equal, the maximum likelihood projection in the column space is the usual orthogonal projection employed for PCA. This conclusion, while not immediately obvious, is intuitively satisfying since there is no

covariance within the columns. By employing Eqs. (33), (34) and (38), the procedure for including correlated errors is greatly simplified and storage requirements are reduced from $m^2 n^2$ to n^2 . The basic algorithm is given in Table 1 and MatLab code for this routine is listed in Appendix A.

3. Experimental

3.1. Data sets

To investigate the potential benefits of incorporating correlated measurement errors into the modeling process using MLPCA, both simulated and experimental data sets were used. The simulated data sets permitted varying degrees of measurement error correlation to be specified under controlled conditions. The experimental data set provided a practical application where the potential benefits of incorporating the correlation of the measurement errors into the modeling process could be examined.

Simulated data sets consisted of spectra from 20 samples of three-component mixtures, with the concentration of each component in each of the 20 mixtures being assigned a value between 0 and 1 from a uniform random number distribution. The spectral profiles of the three components were Gaussian with a σ of 20 nm and maximum molar absorptivities of unity at 480 nm, 500 nm, and 520 nm, respectively. Pure spectral vectors were generated in the range of 400 nm to 600 nm at 5 nm intervals. The noise-free data matrix was calculated by multiplying the 20×3 matrix of concentrations by the 3×41 matrix of pure component spectra. Correlated measurement errors were added to the noise-free data as follows. First, a 20×41 matrix of uncorrelated noise with a uniform standard deviation equal to 1% of the maximum absorbance in the noise-free data matrix was generated from normally distributed random numbers. To introduce correlation among errors within the rows, the rows of this matrix were filtered using a moving average digital filter which was wrapped around at the beginning and end of each row. Filter widths of 5, 7, 9 and 11 points were used in separate runs to examine the effect of the degree of correlation. The correlated noise matrix was added to the noise-free data to

generate the data for calculations. Five replicates for each sample were generated by using the same noise-free data matrix but adding a new noise matrix for each replicate. The replicates were necessary for calculating an ‘observed’ error covariance matrix. Finally, analysis by PCA or MLPCA was carried out using the 20×41 matrix of the means of the five replicates. Calibration models were also built using the same matrix of means and then evaluated using a prediction set of 100 samples with similar characteristics but different concentration values.

The experimental data set was one used previously by Wentzell and Andrews, and the reader is referred to the original work for a complete description of the experiment [5]. Briefly, 26 three-component mixtures of the metal ions Co(II), Cr(III), and Ni(II) were prepared in 4% HNO_3 . Five replicate spectra were obtained for each sample using randomized blocks with a reference spectrum run prior to each

sample to minimize instrumental drift. The spectra were scanned over the range of 350 nm to 650 nm on a HP 8452 diode array spectrophotometer (Hewlett-Packard, Palo Alto, CA) using a standard 1 cm quartz cuvette. Each spectrum was recorded at 2 nm resolution with an integration time of 1 s. In the original work, an optical bandpass filter with a wide pass band was placed in front of the source to decrease its intensity at the edges of the scanned region and therefore increase the level of noise at those wavelengths. The mixture spectra of the samples are shown in Fig. 1.

3.2. Computational aspects

All of the calculations in this work were performed in MatLab (The MathWorks, Natick, MA) on a Sun Microsystems Sparc Server 1000 with 230 MB of memory and four 50 MHz SuperSPARC CPUs.

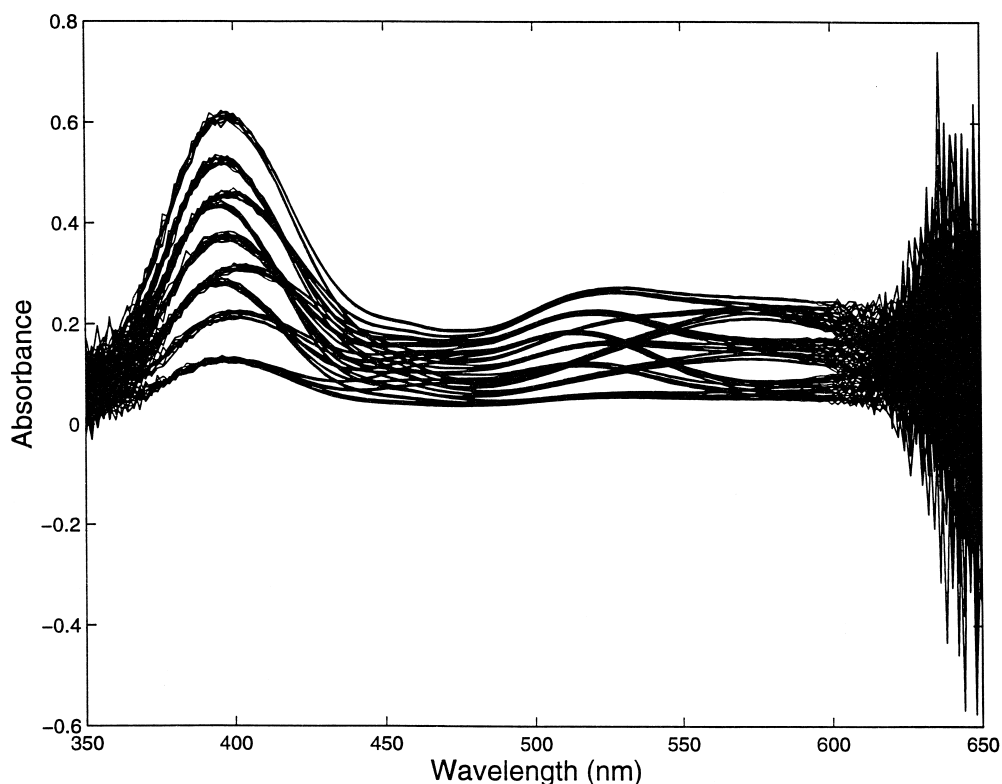


Fig. 1. Spectra of mixtures of cobalt, chromium and nickel ions in 4% HNO_3 .

4. Results and discussion

4.1. Simulated data

In order to distinguish among the various methods for including covariance in the MLPCA decomposition, several cases will be defined. First, the term ‘individual covariance’ will be used to describe those cases where a separate $n \times n$ error covariance matrix is calculated for each sample (row) of the data matrix of mean values. This calculation is made on the basis of q replicate measurements for each sample by Eq. (26). Note that, since the data matrix analyzed is actually the mean of the replicates, the error covariance matrix calculated by Eq. (26) should be scaled by q^{-1} to represent the covariance of the mean, but a constant scaling factor only affects the magnitude of the objective function and not the decomposition results. Individual covariance matrices are used in conjunction with the algorithm described in Section 2.3. A second term that will be used is ‘pooled covariance’. In this case, it is assumed that the error covariance matrix is the same for all rows and an estimate of this can be obtained from Eq. (32) by averaging the individual covariance matrices calculated from the replicates. Finally, the term ‘theoretical covariance’ will be used to describe those cases where true error covariance matrix has been calculated. This is normally only possible with simulated data, but permits a best case reference point for comparison with estimated covariances. In the present study, correlated errors were generated by applying a moving average digital filter to an $m \times n$ matrix of identical and independently distributed (iid) normal errors with a specified variance, \mathbf{E}_{iid} . Mathematically, the matrix of errors added to the data was:

$$\mathbf{E}_{\text{corr}} = \mathbf{E}_{\text{iid}} \mathbf{F}^T \quad (39)$$

In this equation, \mathbf{F} represents the $n \times n$ filter matrix, comprised of the coefficients of the digital filter as a diagonal band and wrapping around at the boundaries. For example, for a five point moving average filter, the n elements of the first row would be (0.2, 0.2, 0.2, 0, ..., 0, 0.2, 0.2), the second row would be (0.2, 0.2, 0.2, 0.2, 0, ..., 0, 0.2), and so on. With this

representation, the theoretical error covariance (sample) is given by:

$$\boldsymbol{\Sigma} = \mathbf{F} \mathbf{F}^T \sigma_{\text{iid}}^2 \quad (40)$$

Here σ_{iid} is the standard deviation of the errors in \mathbf{E}_{iid} .

The first study carried out was intended to see if the algorithm described in Section 2.3 was indeed the same as that in Section 2.4 when the rows of the data matrix had identical covariance matrices. Both theoretical and pooled covariance matrices were used in this study, with five replicates employed for each sample. For the algorithm in Section 2.3, the single covariance matrix was copied m times for all of the samples, while only a single covariance matrix was passed to the algorithm described in Section 2.4. As expected from theory, both methods produced identical results. However, the algorithm in Section 2.4 required less memory and was more efficient by a factor of ca. 500 in this case.

Once the equivalence of the two algorithms for the case of identical error covariance matrices was confirmed, a further study was carried out to examine the effect of including covariance in the data decomposition. To do this, a noise-free 20×41 rank 3 matrix was first generated in the manner described in Section 3.1. This matrix was decomposed by SVD to give ‘true’ values for the decomposition ($\mathbf{U}^0 \mathbf{S}^0 \mathbf{V}^0$). Five separate noise matrices, each with row-correlated errors, were then added to the noise-free data matrix to generate five replicates for each sample. The replicates were used to calculate a mean data matrix and error covariance matrices for each of the 20 rows. Four methods (PCA and MLPCA with individual, pooled, and theoretical error covariance matrices) were then used to decompose the mean data matrix. The resulting vectors ($\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$) were evaluated by calculating the angle with the ‘true’ vectors:

$$\theta = \cos^{-1}(\mathbf{u}_i^T \mathbf{u}_i^0) \text{ or } \cos^{-1}(\mathbf{v}_i^T \mathbf{v}_i^0) \quad (41)$$

This process was repeated with 100 data sets and the mean angles and their standard deviations were calculated over all 100 sets. The results are shown in Fig. 2 for each of the six vectors and four methods of decomposition.

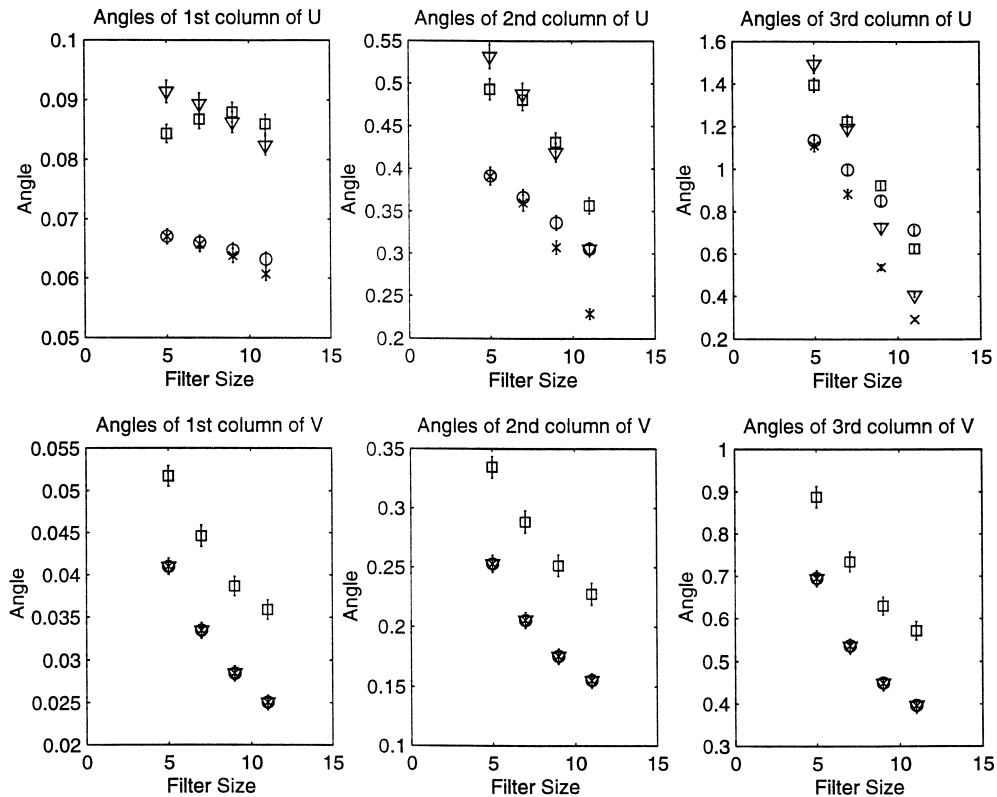


Fig. 2. Comparison of results for the decomposition of simulated data (correlated errors) to the PCA decomposition of error-free data (\circ = PCA, \square = MLPCA with individual covariance estimates, ∇ = MLPCA with pooled covariance estimates, \times = MLPCA with theoretical covariance). Vertical bars show the standard deviation of the means of 100 runs.

The results in Fig. 2 are not intended to be universal in their implications, since the characteristics of this data set are quite specific, but it is instructive to look at patterns of behavior. Focusing on the \mathbf{U} vectors first, it will be noted that use of the theoretical error covariance matrix always produces the best estimate of the true vectors (smallest angle) as might be anticipated. For \mathbf{u}_1 , however, nearly equivalent results are generated with conventional PCA, with deviations becoming more significant with increasing error covariance (larger filter size). The use of individual covariances and pooled covariances gives the poorest results for this vector. This implies that the uncertainty introduced by using estimates of the covariance is more important here than the error introduced by assuming uncorrelated measurement errors. This was confirmed by increasing the number of replicate samples used to estimate the covariance

from 5 to 100, which caused the results from individual and pooled covariances to approach those obtained using the theoretical covariance.

Moving from \mathbf{u}_1 to \mathbf{u}_2 and \mathbf{u}_3 , it is observed that the performance of PCA relative to the theoretical covariance case degrades with the increasing principal component number and increasing error covariance. In contrast, the performance of MLPCA with individual and pooled covariance matrices improves progressively. For the most highly correlated error in the case of \mathbf{u}_3 , PCA performs the worst and the performance of MLPCA with the pooled covariance approaches the best case of the theoretical covariance. Since this is the reverse of the situation for \mathbf{u}_1 , one might ask which is more important. To answer this, the magnitude of the angular deviations can be examined. For \mathbf{u}_1 , the difference in the magnitude of the angular deviations is about 0.02° , while for \mathbf{u}_3 it is

about 0.3° . Since the purpose of PCA is often to estimate the orientation of the hyperplane defined by the vectors (e.g., in target testing), it could be argued that the improvement in the magnitude of the angular deviation observed for \mathbf{u}_3 by including the covariance outweighs the poorer performance observed for \mathbf{u}_1 in this case. Overall, it is clear that a calculation using the theoretical covariance gives the best results, but this is of little practical importance because the true covariance matrix is not generally available. Beyond this, the only general statement that can be made is that the value of including error covariance in the PCA decomposition increases with the reliability of the covariance estimate and the degree of error correlation. The practical utility of including covariance will be demonstrated in Section 4.2.

Turning to the \mathbf{V} vectors, whose performance is evaluated in the bottom half of Fig. 2, a very different pattern is observed. Whenever identical covariance matrices are used in the calculation (PCA and MLPCA with pooled and theoretical error covariance), the performance appears to be identical regardless of the structure of the error covariance matrix. The performance of the calculation with individual covariance matrices is always worse, and once again this was attributed to the uncertainty in the covariance estimates (confirmed by increasing the number of replicates). The seemingly equivalent results obtained with the three other techniques motivated us to examine whether this was the result of a fundamental mathematical equivalence or an accident of symmetry. It can be demonstrated geometrically and numerically that the results are not, in general, mathematically equivalent but are typically very close to one another, especially when the number of rows in the data matrix is large.

It is instructive to try to understand some of the above conclusions from a geometric perspective. To do this, we begin with the simple 2×2 noise-free data matrix given below.

$$\mathbf{X}^0 = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \quad (42)$$

This is clearly a rank 1 matrix and will result in two points on a straight line whether the data are plotted in the row or the column space. Correlated errors were added to these data by first generating iid nor-

mal errors with unit variance and then applying an arbitrary filter matrix in accordance with Eq. (39):

$$\begin{aligned} \mathbf{X} &= \mathbf{X}^0 + \mathbf{E}_{\text{iid}} \mathbf{F}^T = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \\ &+ \begin{bmatrix} 0.5819 & -1.2543 \\ 1.6419 & 1.3455 \end{bmatrix} \begin{bmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{bmatrix} \\ &= \begin{bmatrix} 0.2965 & 2.0310 \\ 4.4344 & 7.5530 \end{bmatrix} \end{aligned} \quad (43)$$

The error covariance matrix for these observations, calculated by Eq. (40) is:

$$\mathbf{\Sigma} = \begin{bmatrix} 0.58 & 0.42 \\ 0.42 & 0.58 \end{bmatrix} \quad (44)$$

This data set was analyzed by PCA and MLPCA with the theoretical covariance matrix. The results, with the rows plotted as two points in the column space, are shown in Fig. 3a. The asterisks ('*') represent the observed data and the line drawn represents the MLPCA solution for \mathbf{v}_1 . The points on this line ('+') are the MLPCA projections of the observed points and the ellipses define the boundaries around these maximum likelihood estimates (contours corresponding to 2σ in the multivariate probability density function). Note that the ellipses are identical (since the row covariances are the same) and are not aligned with the axes (since the errors are correlated). The PCA estimates of the points ('0') are close to, but not exactly on the line defined by the MLPCA solution (the PCA solution for \mathbf{v}_1 would pass through these points, but is not shown for clarity). A line drawn from the origin through the noise-free data points ('x') would represent the 'true' \mathbf{v}_1 . Both the MLPCA and PCA solutions deviate somewhat from the true \mathbf{v}_1 , but they are close to one another. Note, however, that the projections of the points are different for PCA and MLPCA.

The difference between PCA and MLPCA becomes more apparent in Fig. 3b, where the columns of \mathbf{X} are plotted as points in the row space. The same symbols for the points have been used as in Fig. 3a and the line represents the \mathbf{u}_1 vector calculated by MLPCA. In this case, the error boundaries are circular, since there is no correlation between rows and the variances are the same within a column (here they are also the same between columns since homoscedastic

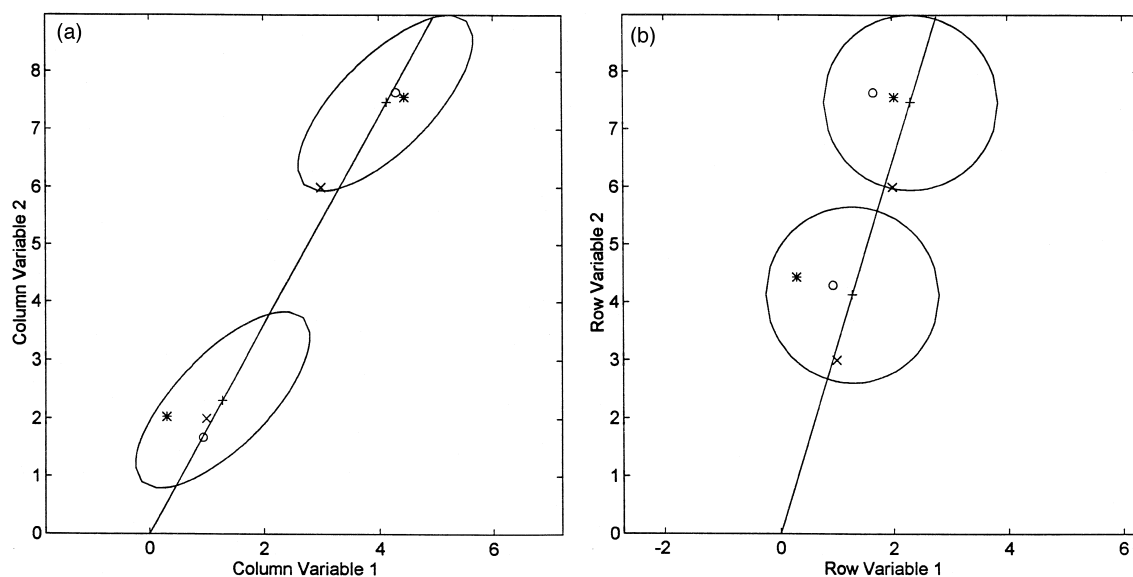


Fig. 3. Geometric illustration of PCA and MLPCA with row error correlations: (a) points plotted in the column space, (b) points plotted in the row space. \times = error-free measurements, $*$ = observed measurements, $+$ = MLPCA projections of the measurements, \circ = PCA projections of the measurements. The line represents the rank one MLPCA solution and the ellipses describe the 2σ boundary defined by the error covariance matrix (centered on the MLPCA projection).

errors were used to generate the correlated errors). The PCA and MLPCA projections are much different in this case, however. This simple example clearly illustrates the differences between MLPCA and PCA when correlated errors are present.

The same noise-free data were also used to illustrate the problem of singular error covariance matrices. As noted earlier, covariance matrices can be rank deficient, leading to problems in their inversion in the calculation of maximum likelihood projection matrices. This rank deficiency can arise either from an insufficient number of replicates in the calculation of an experimental error covariance matrix, or in a theoretical covariance matrix for certain types of digital filters. In this example, the filter matrix was altered to:

$$\mathbf{F} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \quad (45)$$

leading to:

$$\mathbf{X} = \begin{bmatrix} 0.6638 & 1.6638 \\ 4.4937 & 7.4937 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \quad (46)$$

Clearly, the error covariance matrix in this case is rank 1, requiring the numerical adjustments de-

scribed in Section 2.3 prior to inversion. The results of MLPCA in this case are illustrated in Fig. 4. In the column space, shown in Fig. 4a, the geometric interpretation of the singular covariance matrix is evident: the error bounds are confined to a single dimension that defines the direction of the maximum likelihood projection. As in the case of Fig. 3a, the first eigenvector produced by PCA (obtained by connecting the two circles) is similar but not identical to that produced by MLPCA, and both deviate somewhat from the 'true' eigenvector. In the row space, however, the MLPCA solution is much closer to the true solution than the PCA solution, as illustrated in Fig. 4b.

There are a number of implications from the preceding discussion. First, it can be said that when PCA is used to analyze spectra which have correlated errors within each spectrum (as rows in the data matrix), the estimation of the space containing the component spectra is likely to be close to the solution obtained by MLPCA. However, the projections of the observed measurements into that space is likely to be less reliable, leading to possible errors in calibration. Errors in the prediction step are also more likely for

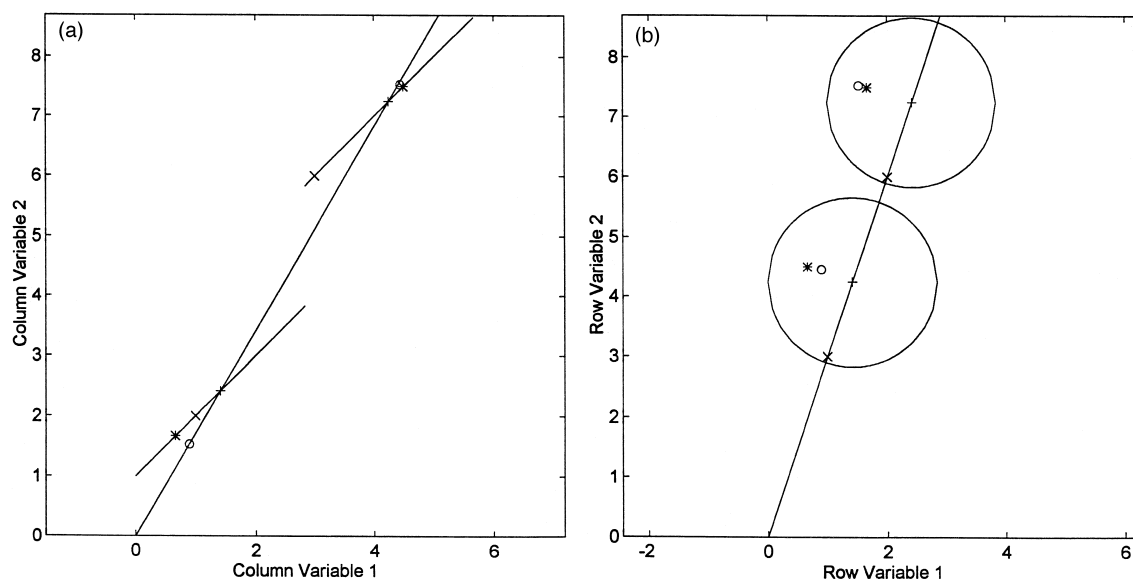


Fig. 4. Geometric illustration of PCA and MLPCA with row error correlations when the covariance matrix is singular: (a) points plotted in the column space, (b) points plotted in the row space. \times = error-free measurements, $*$ = observed measurements, $+$ = MLPCA projections of the measurements, \circ = PCA projections of the measurements. The long line represents the rank one MLPCA solution and the shorter lines describe the 2σ boundary defined by the error covariance matrix (centered on the MLPCA projection).

the same reason. A second implication of the fact that the \mathbf{V} vectors are estimated well by PCA is that the initial approximation to the MLPCA solution is best obtained by carrying out PCA on the original data, and then performing the maximum likelihood projections using Eq. (33). In fact, this has been done for the MatLab code listed in Appendix A. Finally, in cases where an exact solution by MLPCA is too time-consuming, using PCA in conjunction with Eq. (33) is a fast way to obtain an approximate solution which includes covariance. MatLab code for this approach is listed in Appendix B and the viability of this method with experimental data is evaluated in Section 4.2.

To support some of the conclusions drawn above regarding multivariate calibration, principal components regression (PCR) and maximum likelihood PCR (MLPCR) [5] were carried out on the simulated data sets. Each data set consisted of 20 calibration spectra and 100 prediction spectra, each of which was the mean of five replicates. These were generated in the same manner as described earlier. The pooled covariance matrix was calculated from the replicates for the calibration samples. PCR was conducted in the usual way, determining the scores through SVD and re-

gressing the reference concentrations for each of the three components on these scores. In this regression, a pseudorank of three was assumed (expected for the three component mixtures). The calibration model was evaluated using the 100 prediction samples. For each prediction sample, scores were generated through an orthogonal projection of the spectrum onto the first three principal components ($\mathbf{t} = \mathbf{x}\mathbf{V}$, where \mathbf{t} is the 1×3 vector of scores and \mathbf{x} is the $1 \times n$ spectral vector) and these were used to predict the concentration. The root mean squared error of prediction for each component was calculated as:

$$\text{RMSEP} = \sqrt{\sum_{i=1}^{N_{\text{pred}}} (y_i^{\text{pred}} - y_i^{\text{true}})^2 / N_{\text{pred}}} \quad (47)$$

where y_i^{pred} and y_i^{true} are the predicted and actual concentrations of a particular component in prediction sample i and the number of prediction samples is denoted by N_{pred} . A total prediction error was also calculated for each calibration model according to:

$$\text{RMSEP}_{\text{tot}} = \sqrt{\sum_{j=1}^3 (\text{RMSEP}_j)^2 / 3} \quad (48)$$

where $RMSEP_j$ refers to the RMSEP of component j . The lower the $RMSEP_{tot}$ for a calibration model the better the predictive ability of the model. For MLPCR [5], the same basic procedure was used except that MLPCA (using the pooled or theoretical error covariance matrices) was employed in place of SVD and a maximum likelihood projection (Eq. (33)) was used to determine the scores for the prediction samples.

The results of this study are summarized in Fig. 5, which shows the prediction errors as a function of the calibration method and degree of correlation in the errors (filter size). Three calibration methods were used: PCR and MLPCR with pooled and theoretical covariance matrices. For every combination of calibration method and filter size, 100 runs were carried out, each with 20 calibration and 100 prediction samples. The points in the figure show the mean $RMSEP_{tot}$ and the error bars represent the standard deviation of the mean. The results here are consistent

with those in Fig. 2. The smallest prediction errors are always obtained when MLPCR is used with the theoretical covariance matrix. MLPCR with the pooled covariance matrix does not perform as well as PCR when the degree of error correlation is small (filter size = 5), but performs significantly better as the degree of correlation increases. Note that the pattern of behavior is most similar to that for u_3 in Fig. 2, supporting earlier speculation that it is the most poorly defined vector that determines the quality of the regression.

4.2. Experimental data

Although the simulations showed that at least some improvements in the results of multivariate data analysis are possible by incorporating a knowledge of correlated errors through MLPCA, the practical implications remained unclear. For experimental data,

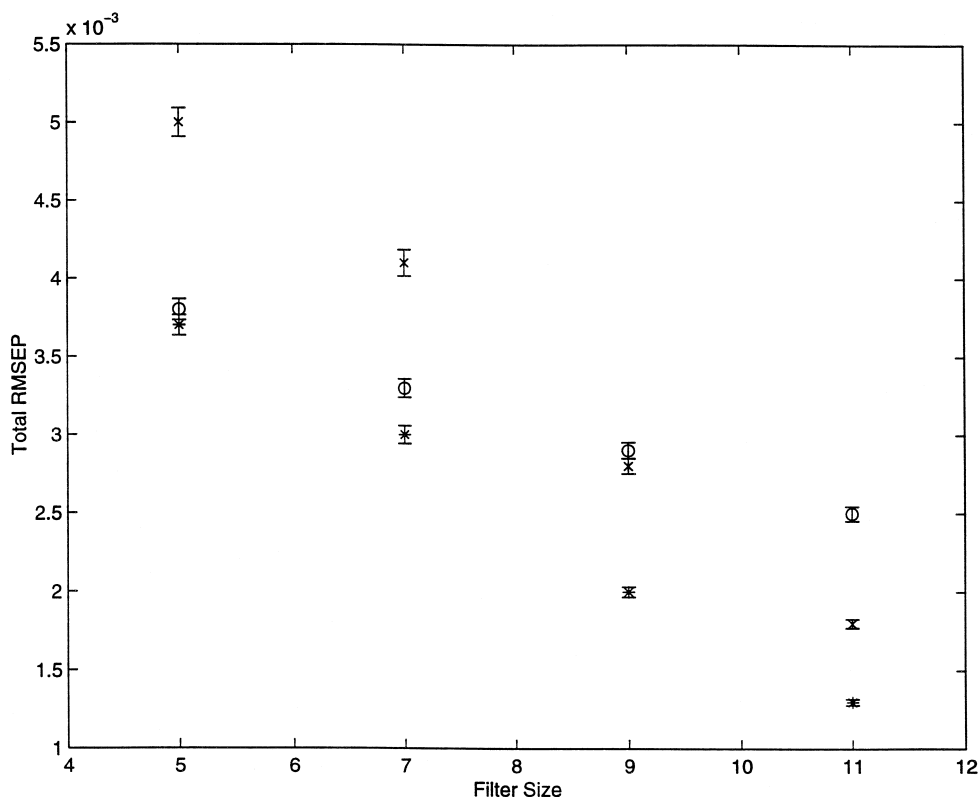


Fig. 5. Results of multivariate calibration by PCR and MLPCR on simulated data sets with correlated errors (○ = PCR, × = MLPCR with pooled covariance estimates, * = MLPCR with theoretical covariance). Vertical bars show the standard deviation of the means of 100 runs.

the exact error covariance matrix is not generally known, and the uncertainty in the pooled covariance estimate may outweigh improvements brought about by its inclusion. Furthermore, the extent of error correlation may be small. To determine if including information about the error covariance matrix in multivariate calibration could be useful in a practical situation, the mixture data from the Co/Cr/Ni system were examined. The residuals from the sample means in this data set were first examined visually and found to be approximately normally distributed. The pooled covariance matrix from this system was then calculated from the replicates and a contour plot of this matrix is shown in Fig. 6. To facilitate visualization, the covariance matrix in the figure was calculated by first dividing the absorbance in each channel by the pooled standard deviation in that channel. This was necessary for the figure because otherwise the large variance in the channels at the edges of the spectrum

dominated the covariance matrix and obscured the structure near the center. For the data analysis, however, the unscaled absorbances were used. The contour plot clearly shows the presence of correlated errors, with particularly high correlations near the center of the spectra and for several channels near the left-hand side. The precise source of these correlations is not known, but it is clear that the assumption of iid errors is not valid here.

Table 2 shows the results of the analysis of this data set using PCR and MLPCR with and without error covariance. In practice, one might be inclined to exclude those regions of the data set that are extremely noisy, so two sets of results are shown. The first includes the full wavelength range, while the second includes only measurements between 388 and 588 nm so that the extremely noisy portions are removed. The results presented are the root mean squared errors of cross-validation (RMSECV) calcu-

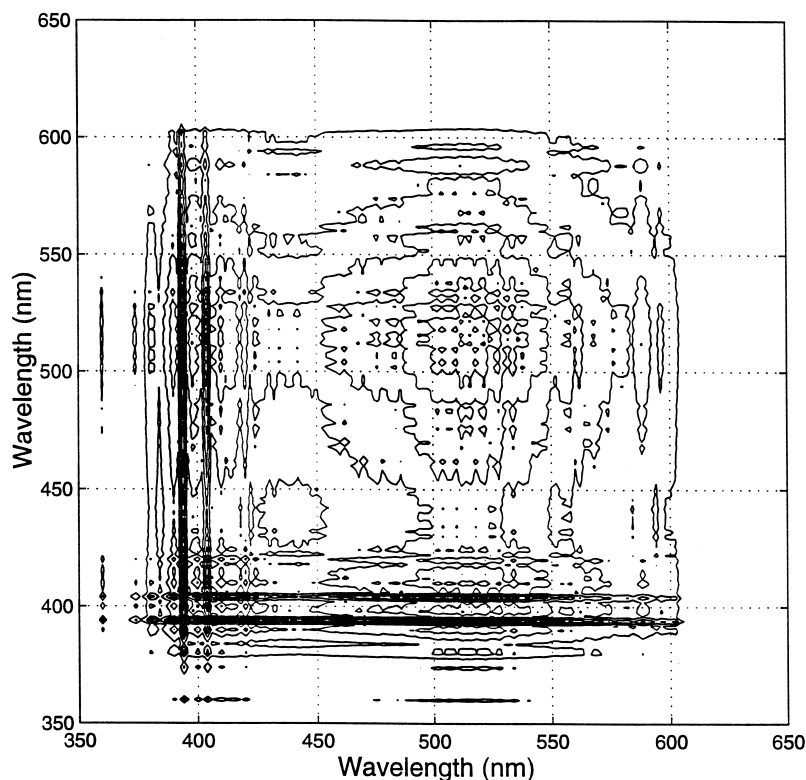


Fig. 6. Contour plot of pooled error covariance matrix (normalized) for experimental data set.

Table 2

Comparison of multivariate calibration results for PCR and MLPCR methods with and without the inclusion of error covariance information (Co/Cr/Ni data set)

Species	RMSECV (mM)					
	Full wavelength range			Reduced wavelength range		
	PCR	MLPCR (no covariance)	MLPCR (covariance)	PCR	MLPCR (no covariance)	MLPCR (covariance)
Co	7.88	0.223	0.141	0.388	0.386	0.188
Cr	2.83	0.076	0.046	0.121	0.114	0.065
Ni	11.02	0.277	0.232	0.433	0.454	0.394
Total	7.99	0.210	0.159	0.343	0.350	0.254

lated through leave-one-out cross-validation. In this procedure, each of the samples in turn is treated as a prediction sample while the calibration is performed on the remaining 25 samples (mean values were used). The RMSECV for a given component is calculated by:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^N (y_i^{\text{pred}} - y_i^{\text{ref}})^2}{N}} \quad (49)$$

where y_i^{pred} and y_i^{ref} are the predicted and reference concentrations, respectively, of a particular component in the excluded sample, and N is the number of samples (26 in this case). A total cross validation error was also calculated for each model in a manner analogous to Eq. (48). Table 2 shows only the results with three latent variables since, as expected, this was the optimum pseudorank. It should be pointed out that the values shown for PCR and MLPCR with no covariance are slightly different from those in the original work [5] because the latter used individual samples rather than mean values.

Considering the results for the full wavelength range first, it is clear that MLPCR, in either form, is far superior to PCR, since it attempts to model the measurement noise. This is consistent with earlier results [5] and will not be discussed further here. The results in Table 2 also show a significant improvement in the RMSECV when the error covariance estimates are included. Although these results are not exactly dramatic, to our knowledge they represent the first case where the use of error covariance information in this way has led to an improvement in results for a practical problem in chemometrics. (Others, such as Næs [10] and Thomas [11] have discussed er-

ror covariance in multivariate calibration, but in a different context.) The extent of the improvement observed in other cases is expected to depend on the degree of correlation of the errors and the reliability with which the error covariance matrix is estimated.

Some interesting observations can also be made for the reduced wavelength range. In this case, the PCR and MLPCR results without error covariance are very similar. This is anticipated, since the error variances are expected to be nearly uniform in this region. In fact, PCR demonstrates a slight advantage, probably because of the uncertainty in estimating the variance for MLPCR. However, a significant improvement is again observed for MLPCR when the covariance is included, indicating the potential importance of this information. It is also important to note that both sets of MLPCR results for the reduced range are inferior to the corresponding results obtained when the full wavelength range is used, and that the best overall results are obtained using the full wavelength range and MLPCR with error covariance. This suggests that, in truncating noisy measurements, we are potentially excluding useful information from the calibration and, by incorporating these measurements in an optimal fashion using an appropriate error model, improved results can be obtained.

It should also be noted that, for this example, identical results were obtained using the fast version of the MLPCA algorithm given in Appendix B, suggesting that the approximation is very reliable. In this case, the results were obtained in a fraction of the time and required only twice as many singular value decompositions as conventional PCR for the cross-validation. Therefore, in the case of identical row er-

ror covariance matrices, arguments against MLPCR based on execution time are not credible and inclusion of the covariance information is actually quite trivial.

5. Conclusions

The principal objective of this paper has been to present more efficient methods to include measurement error covariance in multivariate decomposition by MLPCA. The assumption in developing these methods has been that errors are correlated in the rows of the data matrix only. Further simplifications result if the error covariance matrix can be considered to be identical among all of the rows of the data matrix. Under these conditions, the fastest method for incorporating measurement error covariance is to first decompose the data matrix by SVD, calculate the maximum likelihood projections of the observations using a truncated \mathbf{V} matrix and the error covariance matrix (Eq. (33)), and finally perform SVD on these maximum likelihood projections. However, this algorithm, while fast and compact, is only an approximation to MLPCA (although normally quite a good approximation). The exact MLPCA solution is obtained using the alternating least squares algorithm described in Section 2.4. In cases where the error covariance matrix cannot be assumed to be identical for all of the rows, the algorithm described in Section 2.3 can be used. For all of the methods described here, a solution to the practical problem of a rank deficient error covariance matrix has been proposed and shown to work effectively.

Although an examination of the practical implications of including measurement error covariance information in multivariate analysis was an objective secondary to the theoretical development in this work, some results have been presented which demonstrate the advantages of this approach. Use of the ‘true’ covariance matrix always produced the best results, but in practice the advantages of including error covariance have to be weighed against the uncertainty in estimating the covariance matrix from experimental data and the extent of error correlations. Improved results for one experimental data set have been shown in this work, but further studies with a range of error covariance structures are needed to fully assess the

practical impact of this innovation. Although the effort needed to estimate the error covariance is a barrier to implementation of these methods, this work suggests that perhaps more attention should be paid to understanding the error structure in multivariate data sets so that this information can be more effectively utilized.

Another important area for future work is the development of algorithms for dealing with simultaneous row and column error covariance. These data sets are not uncommon in chemometrics (e.g., chromatography with multichannel detection, fluorescence excitation–emission spectra). While a theoretical foundation has been established to deal with these situations, it is likely to be computationally impractical. Further developments emphasizing numerical simplification could improve this situation.

The work presented here is significant in the establishment of a theoretical and practical framework for accounting for correlated errors in multivariate analysis. It is anticipated that this will be important in many aspects of multivariate analysis, including calibration and mixture analysis, as well as in future studies examining, for example, the effects of digital filtering as a preprocessing tool and the role of highly correlated noise (e.g., baseline drift) in data analysis.

Acknowledgements

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix A

Listing of MatLab code for maximum likelihood principal component analysis with equal row error covariance matrices.

```
function[U,S,V,SOBJ,ErrFlag] = mlcov(X,Cov,p);
%
% MLPCA for equal row covariances.
%
% U,S,V      MLPCA parameters analogous to
%            SVD results
```


$$XX = X * Q * V * \text{inv}(V' * Q * V) * V';$$

$$[U, S, V] = \text{svd}(XX, 0);$$

$$U = U(:, 1:p);$$

$$S = S(1:p, 1:p);$$

$$V = V(:, 1:p);$$

References

- [1] P. Paatero, U. Tapper, *Environmetrics* 5 (1994) 111–126.
- [2] K.R. Gabriel, S. Zamir, *Technometrics* 21 (1979) 489–498.
- [3] H.A.L. Kiers, *Psychometrika* 62 (1997) 251–266.
- [4] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, *J. Chemom.* 11 (1997) 339–366.
- [5] P.D. Wentzell, D.T. Andrews, B.R. Kowalski, *Anal. Chem.* 69 (1997) 2299–2311.
- [6] D.T. Andrews, P.D. Wentzell, *Anal. Chim. Acta* 350 (1997) 341–352.
- [7] J.R. Magnus, H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, New York, 1988.
- [8] H. Martens, T. Næs, *Multivariate Calibration*, Wiley, New York, 1989, p. 230.
- [9] D. Zwillinger (Ed.), *CRC Standard Mathematical Tables and Formulae*, 30th edn., CRC Press, Boca Raton, FL, 1996.
- [10] T. Næs, *Technometrics* 27 (1985) 301–311.
- [11] E.V. Thomas, *Technometrics* 33 (1991) 405–413.