ELSEVIER

Analytica Chimica Acta 350 (1997) 341–352

ANALYTICA
CHIMICA
ACTA

Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer

Darren T. Andrews, Peter D. Wentzell*

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

Received 10 January 1997; received in revised form 11 March 1997; accepted 25 March 1997

Abstract

The application of a new method to the multivariate analysis of incomplete data sets is described. The new method, called maximum likelihood principal component analysis (MLPCA), is analogous to conventional principal component analysis (PCA), but incorporates measurement error variance information in the decomposition of multivariate data. Missing measurements can be handled in a reliable and simple manner by assigning large measurement uncertainties to them. The problem of missing data is pervasive in chemistry, and MLPCA is applied to three sets of experimental data to illustrate its utility. For exploratory data analysis, a data set from the analysis of archeological artifacts is used to show that the principal components extracted by MLPCA retain much of the original information even when a significant number of measurements are missing. Maximum likelihood projections of censored data can often preserve original clusters among the samples and can, through the propagation of error, indicate which samples are likely to be projected erroneously. To demonstrate its utility in modeling applications, MLPCA is also applied in the development of a model for chromatographic retention based on a data set which is only 80% complete. MLPCA can predict missing values and assign error estimates to these points. Finally, the problem of calibration transfer between instruments can be regarded as a missing data problem in which entire spectra are missing on the 'slave' instrument. Using NIR spectra obtained from two instruments, it is shown that spectra on the slave instrument can be predicted from a small subset of calibration transfer samples even if a different wavelength range is employed. Concentration prediction errors obtained by this approach were comparable to cross-validation errors obtained for the slave instrument when all spectra were available.

Keywords: Missing data; Principal component analysis; Calibration transfer; Exploratory data analysis; Maximum likelihood

1. Introduction

Over the past few decades, the development of increasingly more complex analytical instrumentation has demanded the subsequent development of sophisticated models to better extract information from collected data. Of all the different techniques available to analyze multivariate data sets, none is more widely used than principal component analysis (PCA), which has been applied to problems such as mixture analysis, calibration, and exploratory data analysis [1,2]. The strength of PCA stems from its ability to represent multivariate data using a smaller number of variables, called principal components. To do this, the information in many variables is compressed into the first *p*

---

*Corresponding author.

PII S0003-2670(97)00270-5

principal components. If these reproduce the data within experimental error, then p represents the intrinsic rank, or pseudorank, of the data. However, for PCA to be implemented directly, a complete data set (i.e. no missing measurements) is needed, and this is not always the case.

Incomplete data sets commonly arise in a number of situations in chemistry. When modeling chemical data or in exploratory data analysis, it is conceivable that some measurements for a particular sample may not have been recorded or are impossible to obtain experimentally. In other instances, insufficient sample may be available for all measurements, or some measurements may be excluded as erroneous. In the analysis of a time series (i.e. in multivariate statistical process control), measurements may be missing due to sensor failure for a period of time. The problem of calibration transfer, where one wishes to transfer calibration results from a 'master' instrument to a 'slave' instrument based on a small subset of samples, can also be regarded as a missing data problem. In this case, whole spectra from the slave instrument are unavailable. Given the pervasiveness of the missing data problem, it would be extremely useful to have a simple and reliable technique for minimizing the influence of missing measurements in PCA without excluding the incomplete samples entirely. Furthermore, such a method should be have a sound theoretical basis and be capable of predicting missing measurements according to some recognized criterion of optimality.

A number of methods have been developed over the years to handle missing measurements in the multivariate analysis of chemical data [3–8]. Most of these exploit the fact that the intrinsic rank of the data matrix is substantially smaller than the full rank. However, there has been no consensus on a procedure that allows PCA to address this problem in an optimal manner. In this paper, a new approach to the missing data problem is introduced through the application of maximum likelihood principal component analysis (MLPCA), a method described in an earlier paper [9] and recently applied to the problem of multivariate calibration [10]. MLPCA is a generalized form of PCA which allows estimates of measurement uncertainty to be incorporated in the decomposition step. Thus, missing data can be accommodated simply by assigning very large variances to these measurements prior to implementing MLPCA. In this work, the feasibility and advantages of this approach are demonstrated using a number of experimental data sets.

2. Background

As noted above, there have been a number of techniques developed to address the problem of missing data. The simpler of these involve preprocessing of the data prior to the decomposition step in PCA. If the number of missing values is small relative to the size of the data matrix, the usual approach is to discard either all of the samples or all of the sensors with missing measurements (i.e. delete entire rows and/or columns of the matrix). Unfortunately, potentially useful information from the eliminated sensors (or samples) will be lost. As the proportion of missing values increases, this loss of information can be very significant and, in some cases, the pattern of missing data makes this approach impossible. An alternative would be to 'predict' the missing values prior to PCA. The simplest variation of this approach involves the substitution of a zero wherever a missing point is encountered. In certain instances this type of preprocessing may produce acceptable results, but it is not recommended since it can seriously distort the underlying structure of the data. Usually, a better estimate can be obtained via a substitution of the mean of the observed points [11]. Contrary to a popular view, however, substitution of mean values is not neutral from a modeling perspective, and again serious distortion can result.

A more sophisticated approach uses a covariance calculation on the data. Usually, PCA is performed on the original $m \times n$ data matrix, $\mathbf{X}$, using singular value decomposition (SVD) which gives the decomposition,

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}. \qquad (1)$$

Alternatively, the covariance matrix of $\mathbf{X}$ can be decomposed by SVD yielding

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}\mathbf{S}^2\mathbf{V}^{\mathrm{T}} \qquad (2)$$

or

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{U}\mathbf{S}^2\mathbf{U}^{\mathrm{T}}. \qquad (3)$$

If the covariance matrix of $\mathbf{X}$ can be calculated by excluding missing values and reducing the degrees of freedom accordingly, it may be possible to estimate **U** or **V** in this manner. However, the decomposition will weight all of the matrix elements equally and therefore is unlikely to yield an optimum solution. Furthermore, in cases where there are a large number of missing values, this covariance calculation may not be possible.

An extension to the above procedure has been described by Wise in the PLS_Toolbox [12]. In this method, the covariance matrix is calculated and then decomposed by SVD. The intrinsic rank of the data, p, is determined before the analysis and estimates of the missing points are obtained through a regression of the loadings. These values are then substituted into the original data matrix and this updated matrix is used to improve the estimates of the SVD model. This procedure is performed iteratively until either the change in the estimated points falls below a predetermined tolerance or a maximum iteration value is reached. Although this technique proves useful in many instances, it is somewhat cumbersome and its statistical basis has not been explored. Also, if the number of missing values is large, the algorithm may become unstable.

This work investigates the application of MLPCA to the missing data problem. This technique, described recently in the literature [9], performs a PCA-like decomposition of the data but, unlike PCA, uses measurement error variance information to choose the p-eigenvectors. The procedure generates a decomposition which is optimal in a maximum likelihood sense for a model of given dimensionality (p) provided that the measurement errors are distributed as multivariate normal and their variances/covariances are exactly known. In practice, this restriction is seldom met, but that does not detract from the utility of the method (in the same way that the assumptions implicit in linear regression do not limit its use to cases where those assumptions are valid). Usually, independent, normally distributed measurement errors are assumed and sample variances are used. Although these approximations may not be valid, they permit the application of the MLPCA algorithm, which will incorporate the variance information in the decomposition (in contrast to PCA which assumes equal error variances for all measurements). Similar to conventional regression methods, MLPCA minimizes a weighted residual sum of squares which is given by:

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(x_{ij} - \hat{x}_{ij}\right)^2}{\sigma_{ij}^2}, \tag{4}$$

where $\hat{x}_{ij}$ is the maximum likelihood estimate of measurement x_{ij}, and σ_{ij} is the corresponding measurement error standard deviation. The MLPCA decomposition is carried out using a form similar to Eq. (1):

$$\check{\mathbf{X}} = \check{\mathbf{U}}\check{\mathbf{S}}\check{\mathbf{V}}^{\mathrm{T}} \tag{5}$$

where $\check{\mathbf{X}}$ is $m \times n$, $\check{\mathbf{U}}$ is $m \times p$, $\check{\mathbf{S}}$ is $p \times p$ and $\check{\mathbf{V}}^{\mathrm{T}}$ is $p \times n$. It is important to point out here that one of the major differences between PCA and MLPCA is that MLPCA does not have nested solutions (i.e. MLPCA must be performed for each change in the rank estimate, p). Another difference, and a significant advantage, is that the projection of the original data onto the MLPCA eigenvectors is performed using a maximum likelihood projection, which weights the direction of the projection in proportion to the magnitude of the measurement error variances. That is to say, while PCA projections are orthogonal and take the form:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i\tilde{\mathbf{V}}\tilde{\mathbf{V}}^{\mathrm{T}} \tag{6}$$

where $\mathbf{x}_i$ is a row vector of $\mathbf{X}$, and $\tilde{\mathbf{V}}$ is the loading matrix truncated to p principal components, MLPCA projections take the form:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i\Sigma_i^{-1}\check{\mathbf{V}}\left(\check{\mathbf{V}}^{\mathrm{T}}\Sigma_i^{-1}\check{\mathbf{V}}\right)^{-1}\check{\mathbf{V}}^{\mathrm{T}}, \tag{7}$$

where Σ_i is the $n \times n$ covariance matrix for the row vector $\mathbf{x}_i$ (diagonal matrix of variances for independent error). In a projection of this type, less importance will be given to those measurements with large uncertainties. When the errors in all the measurements are the same, Eq. (7) will reduce to Eq. (6). Readers are directed to the original work [9] for further clarification of the differences between PCA and MLPCA.

In this study, the advantages of performing a maximum likelihood decomposition on incomplete data sets will be demonstrated by showing that: (a) even when much of the data is missing, the loadings retain most of the original information; (b) missing data can be reliably predicted; and (c) MLPCA can be used as an alternative approach to calibration transfer and, under the right conditions, produces results similar to those that would be observed if all the spectra on the slave instrument were measured.

3. Experimental

3.1. Data sets

Three experimental data sets were employed to demonstrate the advantages mentioned above. Data Set 1 is a widely used archaeological data set described by Kowalski et al. [13]. These data were obtained in a study intended to relate the origin of obsidian artifacts collected at three different locations to samples taken from four quarries and consist of X-ray fluorescence data on 10 elements for 75 samples. Data Set 2 consists of results compiled by de Ligny et al. [14] in a study of chromatographic retention characteristics. This set contains transformed retention data for 39 solutes with 2 eluents on 6 adsorbents, but is only about 80% complete. Data Set 3 was part of an Infometrix (Seattle, WA) calibration transfer study [15] and consisted of near infrared (NIR) absorbance spectra acquired for 31 samples (mixtures of toluene, chlorobenzene and heptane) measured on two instruments. Further details on each of these data sets are included in Section 4.

3.2. Computational aspects

The calculations in this work were performed using MatLab (The Mathworks, Natick, MA) on two computer platforms: (1) a Pentium-based personal computer and (2) a Sun Microsystems SparcServer 1000 with 230 Mb of memory and four 50 MHz SuperSPARC CPUs. A listing of the MatLab code for the MLPCA routine is given in Appendix A. Under normal circumstances, convergence of the algorithm is quite reliable, although it is considerably slower than for conventional PCA. With missing data, the convergence times are often further extended and may be a problem in extreme cases. Results in this work required times ranging anywhere from under an hour to more than a day. Although this is a drawback to this method, we feel that its sound theoretical basis and versatility justifies its use and are confident that algorithmic improvements (particularly in obtaining initial estimates for the solution) will greatly improve its performance.

4. Results and discussion

4.1. Exploratory data analysis

One of the methods commonly employed to visualize the characteristics of multidimensional data sets is to utilize a projection of the original data into the space described by the first two or three principal components. Samples with similar features often appear 'clustered' together in space after this dimensionality reduction. As an illustration of this approach, we will consider Data Set 1. No pretreatment of the data was performed prior to analysis and, in the absence of other reliable information, uniform measurement errors with a variance of unity were assumed. In the case of a complete data set and uniform errors, the actual magnitude of the error variance is unimportant, since MLPCA is equivalent to PCA under these conditions. In the case of incomplete data, it is only important that the magnitude of variance for missing measurements is much greater than that for the measurements present (a factor of 10^{10} was used in this case).

Fig. 1(a) shows the MLPCA projection of the uncensored data onto the first two eigenvectors. This projection is equivalent to the PCA projection because the measurement error variances are uniform. The grouping of results according to class (1–4 are different quarries; 5–7 are different artifact locations) is consistent with earlier studies and reveals an association between the quarries and the artifacts. However, the actual interpretation of the data is not of interest here, but rather the behavior of the principal components when data are deleted.

To test the ability of MLPCA to deal with missing measurements, 10% of the measurements were eliminated in a systematic fashion. A 'censoring' mask, as shown in Fig. 1(b), was applied to the original data and was generated by removing the first measurement from row 1 (sample 1), the second measurement from row 2, and so on, cycling back to the beginning when the end of each row was reached. Note that, because each sample has at least one measurement missing, this data set cannot be analyzed by conventional PCA

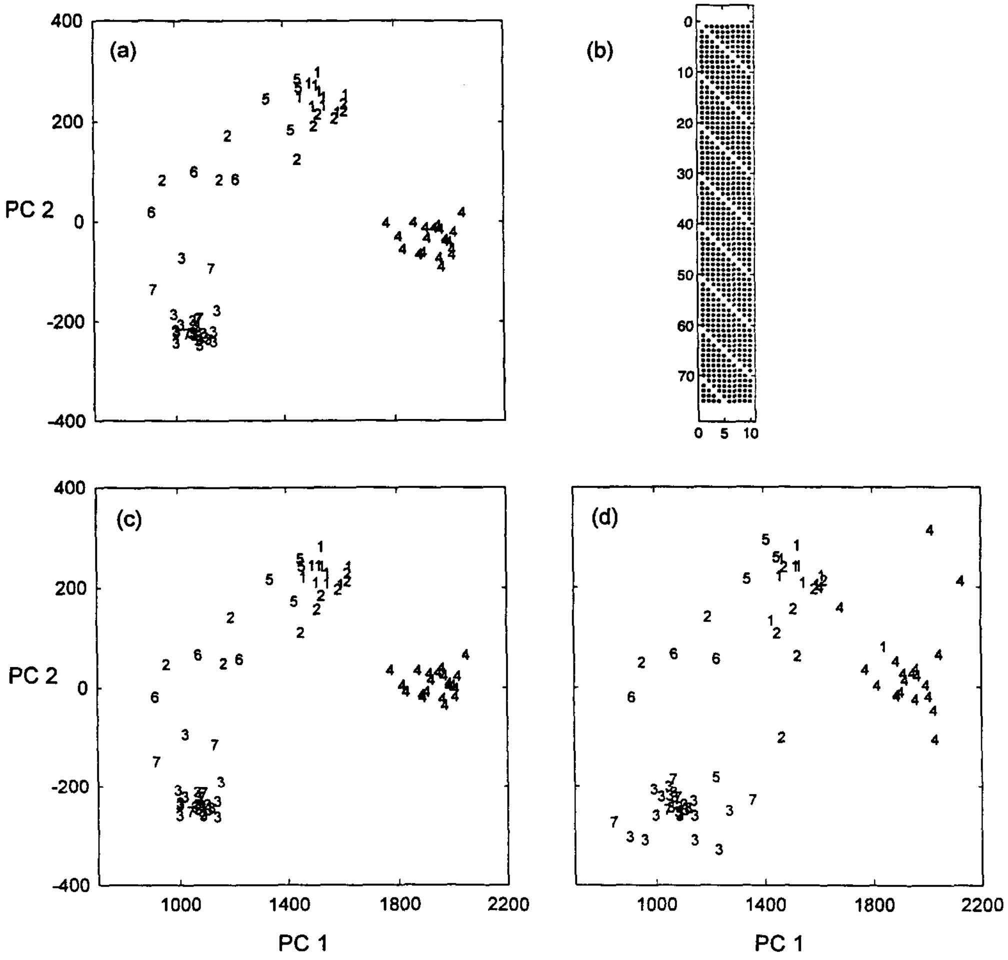

Fig. 1. (a) Projection of archaeological data onto the first two eigenvectors determined by MLPCA. (b) Pictorial representation of data censoring mask for 10% missing data (the white spaces indicate where the data have been removed). (c) Projection of original data into subspace determined by MLPCA for censored data. (d) Maximum likelihood projection of censored data into MLPCA subspace for censored data.

by eliminating rows or columns. However, MLPCA easily handles this situation by employing inflated variances for the missing measurements and the projections are shown in Fig. 1(c) and (d). Fig. 1(c) shows the projection of the original data (i.e. uncensored) onto the first two eigenvectors which were determined by MLPCA performed on the censored data set. This can be regarded as 'cheating' because if the data set were indeed incomplete, the censored measurements would not be available. However, the intent is to demonstrate that even with the missing data, the eigenvectors retain nearly all the information that was contained in the original data. This conclusion is verified by comparison of Fig. 1(a) and 1c, which are virtually identical. Fig. 1(d) is a more accurate portrayal of what one would observe if the measurements were truly absent. In this case, maximum likelihood projections of the censored data onto the MLPCA eigenvectors were used. Although there are small perturbations, it can be seen that the projection very closely resembles the projection for the complete data set in Fig. 1(a), particularly with regard to the clusters observed. Slightly better results were obtained when the data were mean-centered, but this

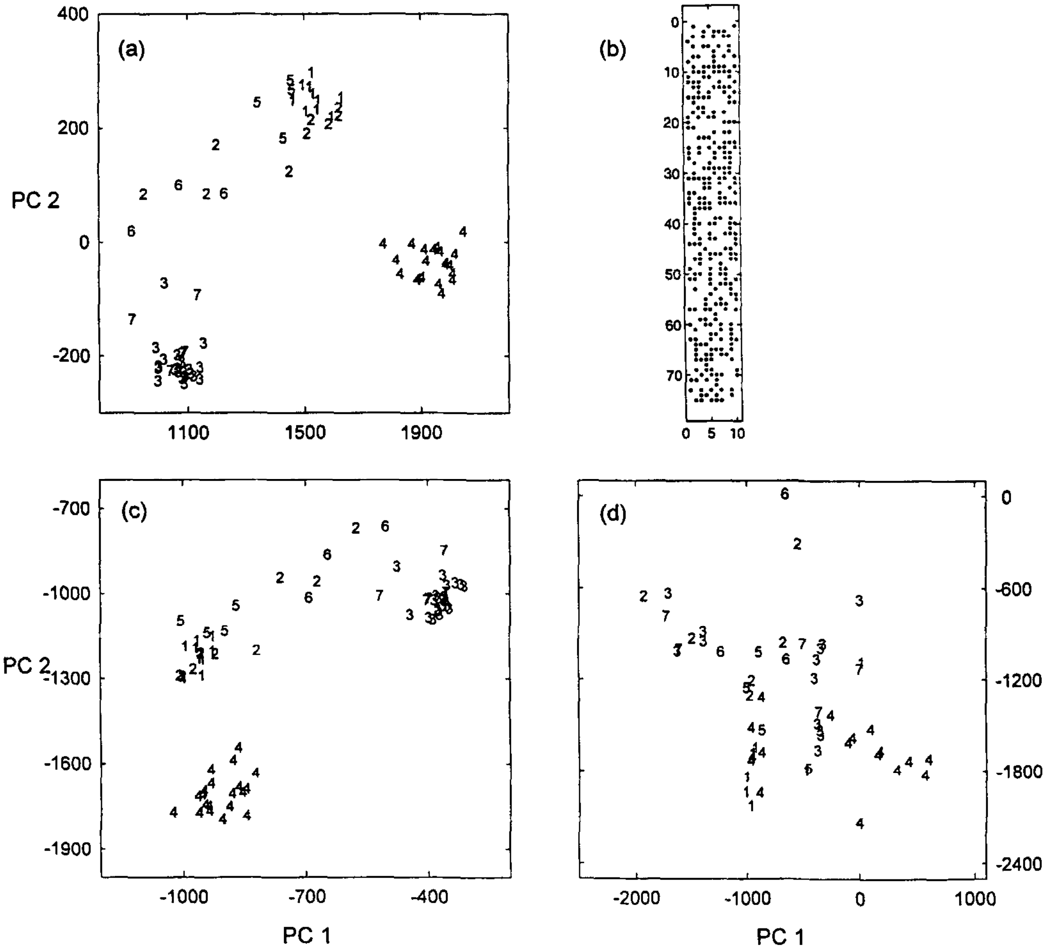

Fig. 2. (a) Projection of archaeological data onto the first two eigenvectors determined by MLPCA. (b) Pictorial representation of data censoring mask for 54% missing data (the white spaces indicate where the data have been removed). (c) Projection of original data into subspace determined by MLPCA for censored data. (d) Maximum likelihood projection of censored data into MLPCA subspace for censored data.

method of preprocessing would not be valid if measurements were truly missing. Of course, the success of such an application will depend on the data set and the correlation of the measurements, but the availability of a statistically optimal procedure for treating missing data is a significant advance.

Data set 1 was also analyzed under more extreme conditions, with as many as 54% of the measurements removed at random. The 'censoring' mask used in this case is presented in Fig. 2(b). Under these circumstances, the representation of the two-dimensional subspace remained quite good. This is illustrated in Fig. 2(c) which shows the projection of the original data onto the eigenvectors for the censored data. Although there is a reorientation of the eigenvectors, the spatial relationship among the samples is largely unchanged. As might be expected, projection of the censored data was not as successful. In this case, the projection is dominated by a few outlying samples which are missing critical measurements. If these are eliminated, the projection in Fig. 2(d) results. Although there is clearly a loss of information in this extreme case, some of the associations are still apparent.

As well as projecting the individual samples into the subspace, a similar treatment may be carried out on the uncertainties associated with the samples. If the $n \times n$ measurement error covariance matrix for a sample (i.e. a row of $\mathbf{X}$) is given by Σ_i, then the $p \times p$ error covariance matrix for the scores of that sample, Ψ_i, will (by propagation of error) be given by

$$\Psi_i = \left(\breve{\mathbf{V}}^{\mathrm{T}} \Sigma_i^{-1} \breve{\mathbf{V}}\right)^{-1}. \tag{8}$$

While this equation does not take into account the uncertainty in the eigenvectors, and so cannot be regarded as a true measure of variance in the scores, it is useful in identifying outliers which are the result of incomplete data. Normally, MLPCA reduces the influence of uncertain measurements through the projection, but in some cases there is insufficient information to do this, resulting in an outlier (e.g. sample 4 in the above example, which only has one measurement). For the archaeological data, a projection of the error covariance matrix for a given sample with uniform, uncorrelated errors will yield a diagonal matrix (2×2) with measurement variances on the diagonal. If the projection of the covariance matrix for some samples with missing data differs from this, the magnitude of the difference will indicate the uncertainty of the scores in the new space. Error bars may be obtained by taking the square roots of the diagonal elements of the error covariance projection and error contours may be constructed utilizing the whole error covariance projection. While caution should be used in interpreting these as true variances of the scores, this approach is useful in the identification of outliers.

4.2. Modeling

The treatment of missing data is expected to be even more successful when strong correlations exist among the measurements. For Data Set 2, retention volumes

available. These data have been analyzed in the past using a physical model proposed by Snyder [16] and three-way factor analysis by de Ligny [14]. The objective in the development of such models is to enable prediction of retention characteristics of solutes with a minimum number of parameters. Estimation of such parameters is made difficult by incomplete data.

Because this is a three-way data set, some preprocessing had to be performed prior to implementation of MLPCA. The easiest manipulation would be to analyze each adsorbent or eluent individually and combine the results. Unfortunately, if much or all of the data are missing for a particular solute in one subset, the predictive power of this technique will be greatly reduced. Therefore, it would be beneficial to use any extra information that may be contained in data for the other adsorbents or eluent in the decomposition step. For this work, the three-way data were 'unfolded' to give a 19×12 matrix where the first 6 columns correspond to the transformed retention data for eluent 1 and the latter 6 for eluent 2.

MLPCA was applied to this unfolded matrix and the simplest model that produced reliable results had a rank of 3. The maximum likelihood projections were then used to reconstruct the missing values, which are listed in Table 1. Also included in the table are the values predicted by de Ligny [14] and Snyder [16]. From the table, it is clear that the MLPCA estimates of the missing data are in excellent agreement with those determined by de Ligny [14] and, with the exception of the last entry, fall within the confidence intervals determined in that work. This agreement is somewhat remarkable given that de Ligny et al use a trilinear model (75 parameters) whereas we have used an unfolded bilinear model (84 parameters). For a comparison, it would be useful to have an estimate of the uncertainty associated with the data predicted by MLPCA. This may be accomplished using an exten-

Table 1
Predictions of missing data for Data Set 2 (transformed chromatographic data)

Adsorbent	Eluent	Sample	MLPCA	de Ligny	Snyder
1	1	5	2.08±0.07	2.22±0.18	2.07
2	1	18	0.53±0.06	0.54±0.13	0.31
3	1	13	0.90±0.03	0.91±0.12	0.53
4	1	18	1.54±0.13	1.53±0.21	1.92
4	1	19	2.29±0.13	2.25±0.23	3.07
5	1	18	1.55±0.12	1.53±0.20	1.94
5	1	19	2.30±0.12	2.26±0.23	3.09
6	1	18	1.61±0.13	1.64±0.22	2.00
6	1	19	2.32±0.12	2.41±0.25	3.12
1	2	7	−0.91±0.12	−0.83±0.18	−0.86
1	2	9	−0.86±0.08	−0.81±0.15	−0.90
1	2	10	−1.05±0.12	−0.96±0.19	−1.18
2	2	8	−0.60±0.05	−0.66±0.13	−1.09
2	2	14	−0.54±0.04	−0.56±0.12	−0.81
2	2	18	−0.41±0.04	−0.42±0.13	−0.86
3	2	1	−0.56±0.07	−0.58±0.16	−0.69
3	2	2	−0.75±0.07	−0.78±0.17	−0.74
3	2	7	−0.70±0.11	−0.69±0.20	−0.67
3	2	8	−0.71±0.08	−0.78±0.17	−0.86
3	2	10	−1.06±0.11	−1.05±0.23	−0.94
3	2	13	−0.45±0.04	−0.45±0.12	−0.30
4	2	1	−0.53±0.08	−0.52±0.17	−0.42
4	2	2	−0.64±0.08	−0.64±0.18	−0.65
4	2	7	−0.52±0.11	−0.49±0.19	−0.48
4	2	8	−0.48±0.09	−0.51±0.18	−0.64
4	2	9	−0.42±0.07	−0.39±0.15	−0.28
4	2	10	−0.99±0.11	−0.94±0.23	−0.54
4	2	18	0.45±0.12	0.46±0.20	0.62
4	2	19	1.07±0.12	1.08±0.23	1.62
5	2	1	−0.45±0.07	−0.50±0.17	−0.39
5	2	2	−0.58±0.07	−0.65±0.18	−0.63
5	2	7	−0.50±0.10	−0.52±0.19	−0.48
5	2	8	−0.46±0.08	−0.54±0.18	−0.65
5	2	9	−0.38±0.06	−0.41±0.15	−0.29
5	2	10	−0.90±0.10	−0.96±0.23	−0.56
5	2	18	0.42±0.11	0.44±0.20	0.61
5	2	19	1.00±0.10	1.07±0.23	1.60
6	2	1	−0.67±0.10	−0.58±0.18	−0.36
6	2	2	−0.75±0.10	−0.67±0.19	−0.51
6	2	7	−0.58±0.13	−0.49±0.20	−0.23
6	2	8	−0.54±0.11	−0.51±0.19	−0.37
6	2	9	−0.48±0.08	−0.40±0.16	−0.02
6	2	10	−1.18±0.13	−1.00±0.24	−0.24
6	2	18	0.54±0.15	0.50±0.22	0.89
6	2	19	1.28±0.14	1.63±0.25	1.88*

the same for all of the data present, a valid estimate of measurement uncertainty is required for use in Eq. (10). De Ligny suggests that an approximate measurement error variance for known values may be obtained using the root-mean-squared error (RMSE) of the fit for the non-missing data. Confidence intervals may then be estimated for data points using:

$$95\%\mathrm{CI} = \hat{x}_{ij} \pm 1.96(\hat{\sigma}_{jj})_i. \qquad (11)$$

In this equation, $\hat{x}_{ij}$ is the predicted measurement and $(\hat{\sigma}_{jj})_i$ is the square root of the jth diagonal element of Σ_i^{pred}. It can be seen from Table 1 that the confidence intervals for the MLPCA prediction are much smaller than their respective counterpart predicted by de Ligny. This results from a better fit of the known data by the MLPCA approach (RMSE =0.041) than that obtained by de Ligny (RMSE =0.10). Although there is a difference in the number of parameters used in the fit, the degree of improvement appears much greater than one would expect.

While the MLPCA and de Ligny results were in good agreement with each other, the estimates by Snyder were often at the limits of the confidence interval of de Ligny and beyond those of the MLPCA model. This suggests that there may be some deficiencies in Snyder's approach. Although MLPCA produced a better fit than de Ligny et al, it cannot be said with certainty which is the best approach since the models are substantially different. However, MLPCA does have certain advantages, namely: (1) it is based on a well-established statistical criterion and is easy to apply; and (2) if error estimates are available for the known data, these can be incorporated into the decomposition step.

4.3. Calibration transfer

With the growing use of multivariate instruments in the workplace, a serious issue that has arisen is that of calibration transfer or instrumental standardization. The problem is that a calibration model determined for a particular instrument may lead to very poor prediction if used with data collected on another instrument. For example, calibration parameters obtained on a laboratory spectrometer may not yield the same results as an instrument used in an industrial setting, even if they are the same instrument model. A variety of sources may give rise to the failure of a calibration model, including variations in bandwidth, noise or sensitivity, differences in wavelength registration, and changes in instrument characteristics with time or the operating environment. One solution to this problem is to rerun all of the calibration samples on each instrument. Unfortunately, this procedure can be very time-consuming, especially if the number of calibration samples and/or instruments is large. Compounding this, environmental changes may require frequent recalibration of an individual instrument, which may in turn require the maintenance of a large number of standards. Therefore, it would be desirable to obtain calibration data on a 'master' instrument and then transfer it to a 'slave' instrument using only the data from a few representative samples. The subject of calibration transfer has been addressed in several recent articles which introduce a variety of techniques [17–20]. An alternative approach, based on MLPCA, is presented here to illustrate the potential of maximum likelihood methods. This study is not intended to be comprehensive, and a more complete analysis and comparison with other calibration transfer methods will be the subject of future work.

The approach to this problem is similar to the missing data examples addressed above, except whole samples from the slave instrument are now regarded as missing. For this work, Data Set 3 is used to illustrate the typical implementation of this technique. The data from the first spectrometer (NIRSystems Model 6500) has dimension 30×201 (A data, shown in Fig. 3(a))

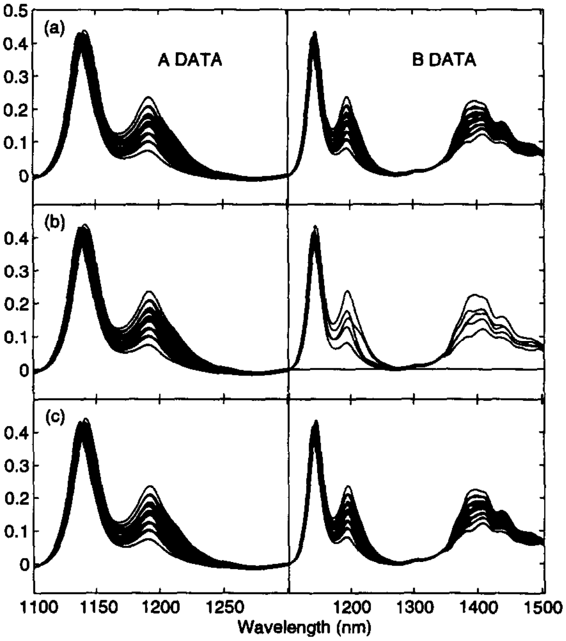

Fig. 3. (a) Plot of master (A) and slave (B) spectra for Data Set 3. (b) Graphical representation of augmented data matrix with 5 representative samples from B. (c) Plot of reconstructed data following analysis by MLPCA.

Appendix

Listing of Matlab Code for MLPCA

```
function [U,S,V,SOBJ,ErrFlag] = mlpca(X,stdX,p);
%
% This function performs MLPCA with missing
% data.
%
% X       is the mxn matrix of observations.
% stdX    is the mxn matrix of standard deviations
%         associated with X (zeros for missing
%         measurements).
% p       is the model dimensionality.
%
% U,S,V   are the pseudo-svd parameters.
% SOBJ    is the value of the objective function.
% ErrFlag indicates exit conditions:
%           0 = normal termination
%           1 = max iterations exceeded.
%
% Initialization
%
convlim=1e-10;
maxiter=50000;
XX=X;
varX=stdX.^2;
[i,j]=find(varX==0);
errmax=max(max(varX));
for k=1:length(i)
  varX(i(k),j(k))=1e10*errmax;
end
n=length(XX(1,:));
%
% Generate initial estimates
%
for i=1:length(X(:,1))
  for j=1:length(X(:,1))
    denom=min([nnz(X(i,:)) nnz(X(j,:))]);
    CV(i,j)=(X(i,:)*X(j,:)')/denom;
  end
end
[U,S,V]=svd(CV,0);
U0=U(:,1:p);
%
% Loop for alternating least squares
%
count=0;
Sold=0;
ErrFlag=-1;
while ErrFlag<0
  count=count+1;
  Sobj=0;
  MLX=zeros(size(XX));
  for i=1:n
    Q=sparse(diag(varX(:,i).^(-1)));
    F=inv(U0'*Q*U0);
    MLX(:,i)=U0*F*U0'*Q*XX(:,i);
    dx=XX(:,i)-MLX(:,i);
    Sobj=Sobj+dx'*Q*dx;
  end
  if rem(count,2)==1
    if (abs(Sold-Sobj)/Sobj)<convlim
      ErrFlag=0;
    elseif count>maxiter
      ErrFlag=1;
    end
  end
  if ErrFlag<0
    Sold=Sobj;
    [U,S,V]=svd(MLX,0);
    XX=XX';
    varX=varX';
    n=length(XX(1,:));
    U0=V(:,1:p);
  end
end
%
% Finished
%
[U,S,V]=svd(MLX,0);
U=U(:,1:p);
S=S(1:p,1:p);
V=V(:,1:p);
SOBJ=Sobj;
```

References

[1] E.R. Malinowski, Factor Analysis in Chemistry, 2nd edn., Wiley, New York, 1991.
[2] K. Beebe and B.R. Kowalski, Anal. Chem., 59 (1987) 1007A.
[3] T.H. Brayden, P.A. Poropatic and J.L. Watanabe, Anal. Chem., 60 (1988) 1154.
[4] G.R. Phillips and J.M. Harris, Anal. Chem., 62 (1990) 2351.
[5] O. Pytella, Collect. Czech. Chem. Commun., 55 (1990) 42.
[6] S. Rannar, P. Geladi, F. Lindgren and S. Wold, J. Chemom., 9 (1995) 459.
[7] C.L. de Ligny, G.H.E. Nieuwdorp, W.K. Brederode, W.E. Hammers and J.C. van Houwelingen, Technomet., 23 (1981) 91.
[8] P.R.C. Nelson, P.A. Taylor and J.F. MacGregor, Chemom. Intell. Lab. Syst., 35 (1996) 45.
[9] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber and B.R. Kowalski, J. Chemom., in press.
[10] P.D. Wentzell, D.T. Andrews and B.R. Kowalski, Anal. Chem., in press.
[11] M.A. Sharaf, D.L. Illman and B.R. Kowalski, Chemometrics, Wiley, New York, 1986.
[12] B. Wise, PLS_Toolbox for use with MATLAB, Version 1.5, Eigenvector Technologies, Manson, WA.
[13] B.R. Kowalski, T.F. Schatzki and F.H. Stross, Anal. Chem., 44 (1972) 2176.
[14] C.L. de Ligny, M.C. Spanjer, J.C. van Houwelingen and H.M. Weesie, J. Chromatogr., 301 (1984) 311.
[15] T. Dean, B.R. Kowalski and R. Pell, Applied Spectroscopy, submitted.
[16] L.R. Snyder, Principles of Adsorption Chromatography, Marcel Dekker, New York, 1968.
[17] T.B. Blank, S.T. Sum, S.D. Brown and S.L. Monfre, Anal. Chem., 68 (1996) 2987.
[18] Y. Wang, D.J. Veltkamp and B.R. Kowalski, Anal. Chem., 63 (1991) 2750.
[19] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni and L. Lazzeri, Chemomet. Intell. Lab. Syst., 27 (1995) 189.
[20] E. Bouveresse, C. Hartmann, D.L. Massart, I.R. Last and K.A. Prebble, Anal. Chem., 68 (1996) 982.
[21] R.W. Kennard and L.A. Stone, Technomet., 11 (1969) 137.