MULTIVARIATE CALIBRATION OF NON-REPLICATED MEASUREMENTS FOR HETEROSCEDASTIC ERRORS

Nirav P. Bhatt, Amit Mitna and Shankar Narasimhan*
Dept. of Chemical Engineering
I.I.T. Madras, Chennai, INDIA 600036

December 24, 2005

Abstract

The quality of multivariate calibration (MVC) models obtained depends on the effective treatment of errors in spectral data. If errors in different absorbance measurements are correlated and have different variances (heteroscedastic), then the Maximum Likelihood Principal Component Regression (MLPCR) method developed by Wentzell et al. [1] is an optimal approach which gives a more accurate MVC model. However, this approach requires either complete knowledge of the error covariances or replicated measurements of all spectra from which an estimate of error covariances can be obtained. We propose a method for developing MVC models from non-replicated measurements when errors in different absorbances are independent, but can have different unknown variances. The core of the proposed approach is an Iterative Principal Component Analysis method which simultaneously estimates the lower dimensional spectral subspace and all the error variances. Application of this approach to simulated and experimental data sets demonstrates that the quality of the model obtained using the proposed method is better than that obtained using PCR, and is comparable to the accuracy of the model obtained using MLPCR.

Keywords: Multivariate Calibration, Principal Component Regression, Maximum Likelihood Principal Component Regression, Heteroscedastic Errors, Errors-in-Variables Regression

1 Introduction

Multivariate calibration (MVC) methods are routinely used for analysis of chemical mixtures. One of the important applications of MVC in chemometrics is the development of a model relating properties of a chemical mixture, such as its composition, to its spectroscopic measurements (absorbances). A variety of methods have been proposed for this purpose. Among them, Principal Components Regression (PCR) is one of the oldest and widely used method.

The development of a multivariate calibration model using PCR is a two-step process. In the first step, Principal Components Analysis (PCA) is used to estimate a lower dimension subspace from the measured absorbance spectra of a set of mixtures. The measured absorbance spectra are projected on to this subspace and their scores are obtained, where the scores are the weights used to represent the projections in terms of the basis chosen for the subspace. In the second step a multivariate linear regression model is developed between the concentrations of the calibration mixtures and their scores. Using this model, the concentration of any new mixture can be estimated from its measured absorbance spectra.

The quality of a MVC model depends on how well the errors in both the spectral and concentration measurements are modelled and taken into account. It is well known that if the errors in different absorbances are assumed to be independent and normally distributed with identical variances (homoscedastic), then

the maximum likelihood estimate of the true subspace is obtained using PCA [2]. Furthermore, in the second step of PCR, the regression model is a maximum likelihood estimate under the assumption that the scores are free of errors and only the concentration measurements contain errors. If errors in different absorbance measurements have different variances and are correlated, then by taking them into account, Wentzell et al. [3] demonstrated that a more accurate estimate of the true subspace can be obtained. The method they developed for this purpose was called Maximum Likelihood PCA (MLPCA). Wentzell et al. [1] used MLPCA as the first step of the PCR algorithm and developed the Maximum Likelihood PCR (MLPCR) method, and showed that significant improvement in the predictive ability of the resulting regression model is obtained.

MLPCA, and consequently MLPCR, requires all error variances and their correlations to be known *a priori*. Typically, this requires replicate measurements of all spectra to be made. From these replicates, an estimate of the error variances and correlations can be directly obtained. In many applications, such replicate measurements may not be available or it may require significant amount of time and resources to perform replicate measurements. It would be advantageous if the error variances and correlations can be estimated simultaneously along with the true subspace from a data set, which does not contain replicate measurements. Recently, Narasimhan and Shah [4] developed a new variant of PCA, known as Iterative Principal Component Analysis (IPCA) which can es-

timate the error covariance matrix and the true data subspace simultaneously. It may be noted, that in the simple case when all errors are independent and have identical variances, then the subspace can be estimated using PCA and the error variance can be estimated from the residuals, without the need for replicate measurements. The IPCA method, on the other hand, can be used to obtain an optimal estimate of the subspace and the error variances, even if the errors are heteroscedastic. IPCA also possesses the following two important theoretical properties.

- It is invariant to any scaling of the data.
- If the dimension of true data subspace is unknown, then it can be exactly estimated by examining the eigenvalues obtained at convergence of the method.

The purpose of this paper is to evaluate the use of IPCA method in developing multivariate calibration models, when errors are heteroscedastic and replicate measurements are unavailable. In particular, we focus on the development of multivariate calibration models for relating concentrations of chemical mixtures with their spectra. The quality of the model developed is evaluated by benchmarking it with the model developed using MLPCR on simulated as well as on an available experimental data set which contains replicated measurements. The results clearly demonstrate, that when replicate measurements are unavailable, then the proposed method can be used to develop a high fidelity

MVC model, which is almost as good as the model developed using MLPCR. The results also clearly show that it is more important to take the variation of error variances along wavelength direction, rather than along mixture direction. Lastly, it is also demonstrated that when replicate measurements are available, the simple technique of using the average of the replicates to develop the MVC model can itself lead to a significant improvement in the quality of the model.

2 Theory

2.1 Problem Formulation

We focus initially on the first step in the development of a multivariate calibration model using PCR, which is concerned with estimating the true data subspace from noisy measurements. Let $y_t(j) : m \times 1$ represent the true values of m variables at the sampling index j. Let these variables be linearly related by p independent equations given by

$$Ay_t(j) = 0 \tag{1}$$

where A: $p \times m$ is a constraint matrix. The rows of A form a basis for a p dimensional subspace of R^n. Equation 1 implies that the true data vectors $y_t(j)$ lie in a $(m - p)$ dimensional subspace of R^n, orthogonal to the row space of A.

At each sampling index j, measurements $y(j)$ of all the variables corrupted by random noises are available which can be written as:

$$y(j) = y_t(j) + \epsilon(j) \tag{2}$$

If we have n such measurement vectors, then we can construct a $m \times n$ matrix Y as

$$Y = [y(1), y(2), \ldots, y(m)] \tag{3}$$

In the case of spectroscopic data, either the rows or columns of Y can represent the spectra of a mixture. For the sake of definiteness, we take each row i of the data matrix to be the absorbance spectra of a mixture i measured at n wavelengths. Then, each column of the data matrix represents the absorbances of m mixtures at a particular wavelength. Typically, the number of mixtures used in developing the calibration model is much less than the number of wavelengths at which the absorbances are measured, and thus the rank of the data matrix is m.

We assume that the random error vectors, $\epsilon(j)$, are mutually independent and follow a multivariate normal distribution with mean zero and covariance matrix Σ_ϵ, that is,

$$E[\epsilon(j)\epsilon^T(k)] = 0 \quad \forall\ j \neq k \tag{4}$$

$$\epsilon(j) \sim \mathcal{N}(0, \Sigma_\epsilon) \tag{5}$$

The random errors are also assumed to be independent of true values of measurements. Properties 4 and 5 imply that the errors can be correlated and can have different variances either along the mixture direction or along the wavelength direction (depending on the interpretation of the measurement vectors) but not both. Such a condition on the error covariance structure is also referred to as

variation along one mode [5]

Given the measured data matrix, the objective is to estimate the $(m - p)$ dimensional subspace of R^n in which the true data vectors lie, and an estimate of the p dimensional subspace of R^n orthogonal to it which corresponds to an estimate of the row space of A. It may also be noted, that regardless of the method used, we will only be able to estimate an arbitrary basis for each of these subspaces. In the following subsections we review how PCA is used to estimate a basis for the true data subspace and use this to motivate the development of our proposed approach.

2.2 Principal Component Analysis

In PCA, an orthonormal basis for the true data subspace is estimated from the orthonormal eigenvectors of the covariance matrix of Y. These orthonormal eigenvectors can be obtained using the truncated singular value decomposition (svd) of the data matrix Y, which is given by.

$$\text{svd}(Y, m) = USV^T \tag{6}$$

where $U : m \times n$ and $V : n \times m$ are orthonormal matrices, and $S : m \times m$ is a diagonal matrix containing the non-zero singular values. It can be shown that the columns of U are normalized eigenvectors of YY^T, the diagonal elements of S are the square root of the eigenvalues, and the columns of V are normalized eigenvectors of Y^TY. Since the rank of Y is m, the number of non-zero singular

values will also be m. Generally, the elements of S are arranged in decreasing order of singular values. Equation 6 can also be written as

$$\text{svd}(Y) = U_1 S_1 V_1^T + U_2 S_2 V_2^T \tag{7}$$

where U_1, S_1 and V_1 correspond to the first $(m-p)$ largest singular values, while U_2, S_2, and V_2 correspond to the smallest p singular values. Then, the columns of U_1 is a basis for the $(m-p)$ dimensional estimated mixture subspace, the columns of V_1 is a basis for the estimated spectral subspace, and the columns of U_2 is a basis for the row space of A, which is orthogonal to the true mixture subspace.

It can be shown that the above estimate is also the solution of the following minimization problem [2].

$$\begin{aligned} \min_{A,\ \hat{y}(j)} \quad & \sum_{j=1}^{n} (y(j) - \hat{y}(j))^2 \\ st \quad & A\hat{y}(j) = 0 \\ & A^T A = I \end{aligned}$$

In the above procedure, it has been tacitly assumed that we know the dimension of the true data subspace. This assumption may be justified for the specific problem we are dealing with. The spectra of a mixture is usually a linear combination of pure component spectra. Therefore, the dimension of the true data subspace, $(m-p)$, will be equal to the number of species s in the mixture, which is known. However, if a common baseline spectra correction (offset) has to be applied to

all mixture spectra, then the dimension of the true spectral subspace will be one more than the number of species. Henceforth, we assume the dimension of the true data subspace to be s.

2.3 Maximum Likelihood Principal Component Analysis

The least squares objective function minimized by PCA is optimal, if the measurement errors in all spectra are normally distributed, independent, and have identical variance. But in practice this assumption may not be valid due to variations in source intensity or variations in detector noise characteristics, etc. Wentzell et al. [3] have developed a new method called maximum likelihood principal component analysis (MLPCA) for estimation of true data subspace, that can take into account measurement errors which are correlated and whose standard deviations vary with from mixture to mixture as well as from wavelength to wavelength. We are, however, only restricting our considerations to the case when the error covariance structure varies along only one mode (either mixture or wavelength direction). Under this restriction, it can be shown that the objective function which is minimized in MLPCA is a weighted sum square of residuals given by

$$S^2 = \sum_{j=1}^{n} (y(j) - \hat{y}(j)) \Sigma_{\epsilon}^{-1} (y(j) - \hat{y}(j)) \tag{8}$$

The estimate vector $\hat{y}(j)$ must also satisfy Eq. 1. Minimizing the above objective can be shown to be identical to maximizing the likelihood function of the

measurements.

If the covariance matrix Σ_ϵ is known, then the optimal estimate of the true data subspace can be obtained by applying PCA to transformed measurements as follows [4, 5]. Let the cholesky factorization of the covariance matrix be given by

$$\Sigma_\epsilon = LL^T \tag{9}$$

The transformed data matrix is defined by

$$Y_s = L^{-1}Y \tag{10}$$

The truncated svd of the transformed data matrix Y_s can be written as

$$\text{svd}(Y_s, m) = U_{1s}S_{1s}V_{1s} + U_{2s}S_{2s}V_{2s} \tag{11}$$

where U_{1s}, S_{1s}, V_{1s} corresponds to the first $(m-p)$ largest singular values. The columns of U_{1s} represents a basis for the subspace of transformed true data vectors. A basis for the true data subspace can now be obtained as

$$\hat{B} = LU_{1s} \tag{12}$$

and a basis for the row space of constraints is estimated as

$$\hat{A} = U_{2s}^T L^{-1} \tag{13}$$

It may be noted that the estimated bases (columns of B and rows of $\hat{A}$) are not orthonormal. An interesting property of the above procedure is that the expected

value of the largest $(m-p)$ singular values are strictly greater than unity, while the expected value of the remaining singular values are exactly equal to unity [6]. This property can be used to choose the correct dimension of the true data subspace, in case it is not known *a priori* [4].

Wentzell et al. [3] proposed iterative algorithms which embed PCA in an alternative regression technique to obtain the true data subspace. These methods are also applicable to general noise covariance structures. However, like the above approach complete information regarding the error covariances must be known *a priori* in order to apply them.

An estimate of the error covariances can be obtained by replicate measurements. If such replicates are available and can be obtained without expending too much effort or resources, then these methods are very useful and can lead to significant improvement in the predictive abilities of the multivariate calibration models. However, where such replicate measurements are unavailable and are expensive to obtain, it would be advantageous to develop a method that can simultaneously obtain both an estimate of the true data subspace and error covariances from non-replicated measurements. Such a method referred to as the iterative PCA (IPCA) method has been recently developed by Narasimhan and Shah [4] under some additional restrictions. In the following section we provide a brief description of this method.

2.4 Iterative Principal Component Analysis

The IPCA method is applicable only to the special error covariance structures that obey the properties given by Eqs. 4 and 5. In other words, the error covariance can vary along only one mode (either mixture or wavelength direction, but not both). Furthermore, it is assumed that the positions of the zero and non-zero elements of the error covariance matrix Σ_ϵ are known, even though their values are not known *a priori*.

IPCA estimates the non-zero elements of the error covariance matrix and a basis for true data subspace simultaneously, using the following procedure.

- **STEP 1**. Assume that an initial estimate $\hat{A}^0$ of the constraint matrix A is available (such an initial estimate can be obtained using PCA). Using the initial estimate, the constraint residuals for each sample j can be computed as

$$r(j) = \hat{A}^0 y(j) \tag{14}$$

  Under the assumption made for the measurement error vectors, these constraint residual vectors can be shown to be independently and identically distributed Gaussian variables with zero mean and covariance matrix $\Sigma_r = \hat{A}^0 \Sigma_\epsilon (\hat{A}^0)^T$. The joint distribution of the residual vectors can thus be obtained and an estimate of the non-zero elements of Σ_ϵ can be obtained by maximizing, the logarithmic likelihood function of these residuals. The

resulting optimization problem is formulated as :

$$\min_{\Sigma_\epsilon} \quad n \log |\hat{A}^0 \Sigma_\epsilon (\hat{A}^0)^T| + \sum_{j=1}^{n} r^T(j) (\hat{A}^0 \Sigma_\epsilon (\hat{A}^0)^T)^{-1} r(j) \tag{15}$$

The above optimization function is used to get an estimate of error covariance matrix. The optimization of 15 can be carried out using a nonlinear optimization technique. Constraints can be imposed to ensure that the estimated error covariance matrix is positive definite.

- **STEP 2**. Let the estimate of Σ_ϵ obtained in STEP 1 be denoted as $\hat{\Sigma}_\epsilon^1$. If we denote the cholesky factor of the estimated error covariance matrix as L^1, then

$$\hat{\Sigma}_\epsilon^1 = L^1 (L^1)^T \tag{16}$$

It was pointed out in the preceding subsection that if the error covariance matrix is known, then the true data subspace can be estimated by applying PCA to the transformed data matrix. Following 10, the cholesky factor of the estimated error covariance matrix is used. The transformed data matrix Y_w^k at iteration k is given by

$$Y_w^k = (L^k)^{-1} Y \tag{17}$$

The above transformation is equivalent to scaling the data using standard deviations of the corresponding measurement errors in the special case of a diagonal error covariance matrix. By performing an svd of the weighted

data matrix we get a new estimate of the constraint matrix

$$svd(Y_w^k, m) = U_{1w}^k S_{1w}^k V_{1w}^{kT} + U_{2w}^k S_{2w}^k V_{2w}^{kT} \tag{18}$$

where as before U_{1w}^k, S_{1w}^k, and V_{1w}^k correspond to the first $(m-p)$ largest singular values of the transformed data matrix Y_w^k. The estimated constraint matrix in the transformed domain $\hat{A}_w^k$ is given by the transpose of U_{2w}^k. From this the estimated constraint matrix in the original variable space can be obtained as

$$\hat{A}^k = \hat{A}_w^k (L^k)^{-1} \tag{19}$$

Using the estimated constraint matrix $\hat{A}^k$ at iteration k, the estimate of $\hat{\Sigma}_\epsilon^{k+1}$ is obtained by performing the minimization as in Step 1 described above.

The two steps of the above procedure are repeated until the estimates of $\hat{\Sigma}_\epsilon^k$ and $\hat{A}^k$ converge. A simple test of convergence is to check that the singular values obtained using PCA do not change significantly from one iteration to the next.

It should be noted, that since the non-zero elements of Σ_ϵ are estimated using the covariance information of the constraint residuals, the number of elements of the error covariance matrix that can be estimated depends on the rank of the sample covariance matrix of constraints residuals, AYY^TA^T. In the case when the measurement error covariance matrix is constant with respect to wavelengths, then the rank of the YY^T is m and rank of A is $p = m - s$. Thus the rank of

the sample AYY^TA^T is $(m-s)$. On the other hand, if we assume that the error variances vary with respect to wavelengths and constant for all mixtures, then we have to transpose matrix Y and apply PCA, MLPCA or IPCA to estimate the true data subspace. In this case, the rank of Y^TY is still m, while the rank of A is $p = n - s$. The rank of AY^TYA^T is the minimum of m and $(n - s)$. Since n is usually much greater than m, the rank of AY^TYA^T is m. For generality, we denote the rank of the sample covariance matrix of constraints residuals by r_V. The maximum number of diagonal and non-zero off-diagonal elements of Σ_ϵ that can be estimated is $r_V(r_V + 1)/2$. This condition is not a limitation of the IPCA method but a necessary condition for simultaneous estimation of error covariances and a basis for the true data subspace from non-replicate measurements. In fact, the problem we are considering here is identical to the functional regression problem for errors-in-variables model that is well studied in statistics literature [7], and the above restriction corresponds to the identifiability condition imposed for solving this problem. It should also be noted that the estimates of Σ_ϵ and the basis for the true data subspace are not maximum likelihood estimates of the measured data. It has been proved that even for the simple bivariate case, maximum likelihood estimation procedure breaks down for simultaneous estimation of error variances and regression parameter for the errors-in-variables functional regression problem [8].

The above procedure can be applied for estimating the error covariance matrix

for mixtures assuming that it is constant for all wavelengths, or for estimating the error covariance matrix for wavelengths assuming that it is constant for all mixtures, subject to the identifiability condition being satisfied.

As pointed out in the preceding subsection, if the dimension of the true data subspace is chosen correctly and the estimated error covariance matrix converges to the true covariance matrix, then only the first s singular values of the transformed data will be greater than unity, while the remaining non-zero singular values will all be equal to unity. This feature can be exploited to precisely estimate or check the dimension of the data subspace chosen [4].

2.5 Development of MVC model

Once a basis for the true data subspace is estimated, the second step is to obtain the regression model relating concentration of the mixtures to their true absorbance spectra. An estimate of the true absorbance spectra of mixtures can be obtained by projecting all the measured absorbances on to the estimated spectral subspace. These projected vectors can be represented in terms of the basis for the spectral subspace. The coefficients of this representation are also known as scores. A linear regression model relating measured concentrations and the scores of the calibration mixtures is the desired MVC model.

In the case of PCA, orthogonal projections give the best estimates of the true data vectors corresponding to each measured spectra. The corresponding scores

matrix is given by

$$T_{PCA} = YV_1 \tag{20}$$

In the case of MLPCA, we first need to obtain the maximum likelihood projections (estimates) of the measured spectra and then represent them in terms of the basis for the spectral subspace in order to obtain the scores. The maximum likelihood estimates ($\hat{Y}_{MLE}$) of the measured data are obtained as part of the alternating regression algorithm developed by Wentzell et al. [3].

If it is assumed that the error variances vary only with respect to mixtures, then the maximum likelihood estimates of the measured data can be obtained in IPCA as

$$\hat{Y}_{MLE} = Y[I - \hat{\Sigma}_\epsilon \hat{A}^T (A\Sigma_\epsilon A^T)^{-1} \hat{A}]Y \tag{21}$$

where $\hat{\Sigma}_\epsilon$ and $\hat{A}$ are the converged estimates obtained.

From the svd of $\hat{Y}_{MLE}$ an orthonormal basis for the estimated mixture and spectral subspaces can be obtained.

$$\text{svd}(\hat{Y}_{MLE}, s) = U_1 S_1 V_1^T \tag{22}$$

The scores are computed as in 20 by

$$T_{MLE} = \hat{Y}_{MLE} V_1 \tag{23}$$

The linear regression model relating concentrations to the scores can be written as

$$C = TM \tag{24}$$

where $C : m \times s$ is the matrix of measured concentrations of s species in the m calibration mixtures. Then the ordinary least squares solution for the regression matrix M is given by

$$M = (T^T T)^{-1} T^T C \tag{25}$$

A further point has to be noted when IPCA is applied to data under the assumption that error variances vary with respect to wavelengths and not with respect to mixtures. In this case, if the data matrix is defined as in Eq. 3, then it has to be transposed before applying the IPCA algorithm to estimate the error covariance matrix as well as the true data subspace. The maximum likelihood estimates obtained using Eq. 21, has to be transposed before determining its singular value decomposition.

The names PCR, MLPCR, and IPCR are used to denote the methods in which PCA, MLPCA, and IPCA are, respectively, used as the first step of the MVC model development.

2.6 Prediction using MVC model

For predicting the concentrations of a new mixture given its absorbance spectra $y_{new} : 1 \times n$, the approach used in PCR is to obtain the scores for the new mixture by orthogonal projection followed by the use of the regression matrix. The equation for this purpose is given by

$$t_{new} = y_{new} V_1 \tag{26}$$

and the concentrations are predicted as

$$c_{new} = t_{new} M \tag{27}$$

If the error variances do not vary with respect to wavelength, then the above equations can be used to predict the concentrations of the new sample in the case of MLPCR or IPCR, with the understanding that the orthonormal basis matrix V_1 is obtained using Eq. 22.

However, if the error variances are assumed to vary with respect to wavelength, then the scores for the new sample in the case of MLPCR have to be obtained using the maximum likelihood projection.

$$t_{new} = y_{new} \Sigma_{new}^{-1} U_1 (U_1^T \Sigma_{new}^{-1} U_1)^{-1} \tag{28}$$

where Σ_{new} is the specified error covariance matrix for the new mixture. The concentrations are predicted using Eq. 27. Similarly in IPCR, the scores for the new sample are obtained using MLE projection, except that the estimated covariance matrix $\hat{\Sigma}_\epsilon$ is used instead of Σ_{new} in Eq. 28, since it is assumed that the error covariance matrix of the mixture is unknown, but is identical for all mixtures.

3 Description of data sets

3.1 Simulated data sets:

The quality of MVC models developed using the proposed IPCR method is evaluated by applying it to simulated absorbance data as well as to an available

experimental UV data set of mixtures of three species. Both PCR and MLPCR were also applied to the same data sets for comparative evaluation.

Simulated absorbance data for mixtures of three hypothetical species were generated following the procedure described by [3] as follows:

1. The spectral profiles of the three species were taken to be Gaussian distributions, with a peak absorbance of unity at 480nm, 500nm, and 520nm, respectively, and a standard deviation of 20nm. Pure component spectral vectors were generated between 400 nm and 600 nm at intervals of 5 nm to obtain a 3×41 pure component spectral matrix.

2. Concentrations of twenty mixtures (assumed to be in millimolar units) of the above three species were generated by choosing random numbers between 0 and 1 from a uniform distribution for each species. We thus obtain a 20×3 three-species mixture concentration matrix.

3. The true (or noise free) data matrix of 20×41 was calculated by multiplying the mixture concentration matrix by pure component spectra matrix.

In order to simulate noisy measurements, random errors were added to the above true absorbance matrix according to the specification of the error covariances. Three different data sets corresponding to different error covariance structures are used to obtain a fair comparison between the different methods. In data set 1, the error covariance matrix corresponding to all wavelengths was

assumed to be diagonal, with different variances for different wavelengths. This covariance matrix is assumed to be the same for all mixtures. This case corresponds to variation of error variances along wavelength direction only. Data set 2, was generated by assuming the errors in different absorbance measurements to be mutually independent, but having different variances. This corresponds to the case when the error variances vary along both wavelength and mixture directions. However, the error covariance structure in either of these directions is a diagonal matrix. The third data set was generated by assuming the errors in different absorbances to be correlated and having different variances. This corresponds to the case when the error variances vary along both directions. Moreover, the error covariance structure in either of these directions is non-diagonal. for all three data sets, the measurement errors are generated in two steps. All three measured data sets are of size 20×41.

For data set 1, the standard deviation of the error in absorbance at a particular wavelength was taken to be 5% of the maximum of the true absorbances among all mixtures at the corresponding wavelength. A 20×1 random vector is generated from a $N(0, 1)$ distribution and multiplied with this standard deviation. This is repeated for all wavelengths and the data arranged to get the 20×41 error matrix, which is added to the true data matrix to obtain the measured absorbances.

The standard deviation of error in measurement of absorbance of a mixture

at a particular wavelength is taken to be equal to 5 percent of corresponding true absorbance, for generating data set 2. A random number from an $N(0,1)$ distribution is multiplied with this standard deviation to obtain the error and added to the true absorbance to obtain the measured absorbance. This is repeated for all mixtures and absorbances.

Data set 3 contains measured absorbances containing errors which are correlated and whose variances also vary both with respect to wavelength and mixture. For simplicity, the error in the absorbance of a mixture at a particular wavelength is correlated only with a few of the other errors and not with all the rest. The procedure we used is identical to that used by Wentzell et al., (1997a) for generating their simulated data set 7. For this purpose, first an uncorrelated error matrix E (20×41) was generated as in the case of data set 2. In order to generate the error in the first mixture at the first wavelength, ϵ_{11}, we use a filter matrix Φ_{11} of size 20×41 with non-zero elements as indicated below.

$$\Phi_{11} = \begin{bmatrix} \bullet & \bullet & & & \bullet \\ \bullet & \bullet & & & \bullet \\ & & & & \\ \bullet & \bullet & & & \bullet \end{bmatrix} \tag{29}$$

The non-zero elements of the filter matrix are all taken to be equal to 1/9. The error ϵ_{11} is obtained as

$$\epsilon_{11} = vec(\Phi_{11})^T vec(E) \tag{30}$$

where vec(.) is a vector obtained by stacking the columns of a matrix one below the other. The filter matrix Φ_{ij} for generating the error in the absorbance of

mixture i and wavelength j is obtained by shifting the rows Φ_{11} down by $i-1$ rows and columns of Φ_{11} to the right by $j-1$ columns with wraparound. The errors generated using Eq. 30 in this manner is added to the true absorbances to get the noisy data matrix.

Data set 4 contain absorbances of mixtures of three metal ions (Co(II), Cr(III), and Ni(II)) prepared in 4 per cent HNO_3 solution, obtained through carefully designed experiments by Wentzell et al. (1997a). The data set contains absorbance measurements for 26 mixtures between the range of 300nm-650nm at intervals of 2 nm. Corresponding to each mixture, five replicate measurements of its absorbance spectra have been made, each of which have been given as a separate spectra in the data set. Near the ends of the wavelength range, the noise levels were increased by using a band-pass filter and thus, the variance of the noise varies with wavelength for this data set. The standard deviations of errors at different wavelengths can be estimated directly using the five replicates. The spectra and standard deviations directly estimated from the replicate experiments for this data set are shown in figures 1 and 2.

4 Results and Discussion

4.1 Comparison Methodology

The predictive ability of the calibration models constructed by different methods were validated using the leave one score out cross-validation approach described in [1]. In this approach, the true data subspace is estimated using all the mix-

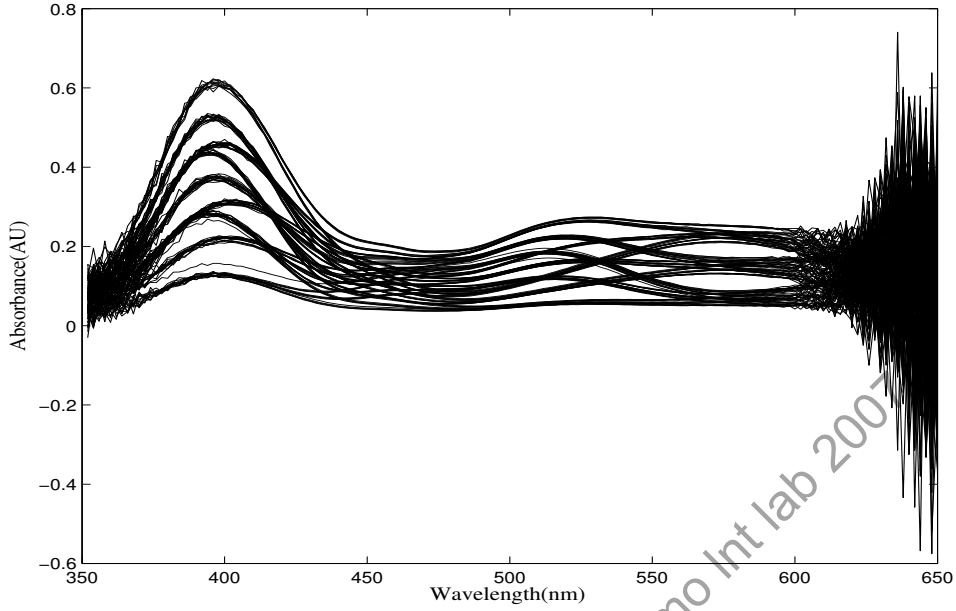

Figure 1: Spectra for metal ion mixtures (UV data set)

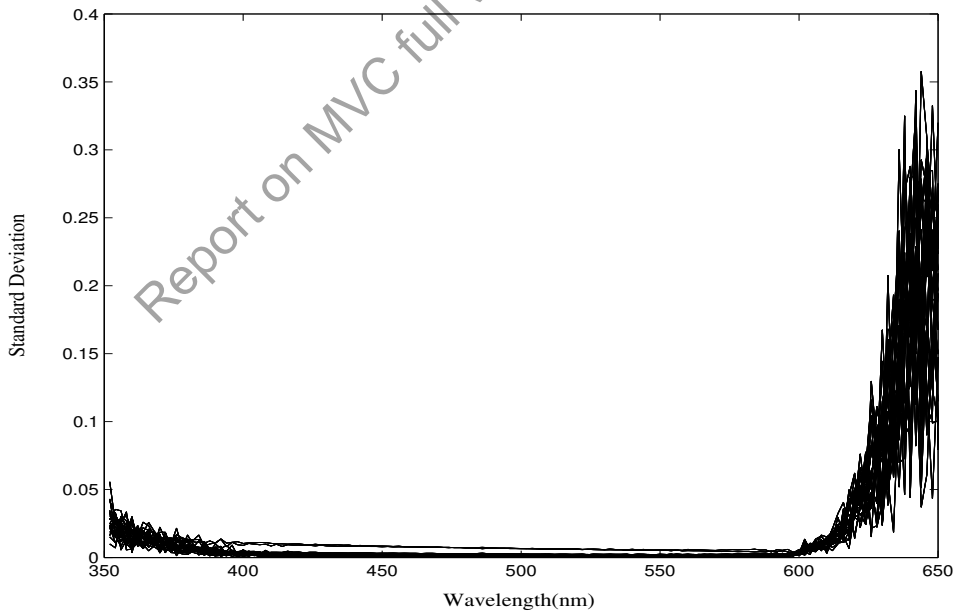

Figure 2: Standard deviations of measurement errors for experimental data set

tures, but the calibration model is developed by excluding the scores of one of the mixtures. The MVC model is used to predict the concentration of the mixture whose scores were excluded. This process is repeated such that each sample is excluded once, and the root-mean square error (RMSE) between predicted and actual concentrations is calculated by

$$RMSE = \sqrt{\sum_{i=1}^{m}(c_i^{pred} - c_i^{ref})^2/m} \tag{31}$$

where c_i^{pred} and c_i^{ref} are the predicted and reference concentrations, respectively, of the species in the excluded mixture. An overall or total root-mean square error total (RMSET) can be calculated as square root of the sum of square of RMSE for all the substances present in the mixture. These values give an indication of the predictive ability of the model.

4.2 Performance Comparison on Simulated Data Sets

We first present the performance of PCR, IPCR and MLPCR on simulated data sets. The RMSE results for data set 1 are presented in Table 1. Since the number of species in the mixtures is three, the dimension of the true data subspace is also three. Nevertheless, we present the results obtained by assuming different values for the dimension of the true data subspace. It is observed from Table 1, the lowest RMSET values for all methods are obtained when the dimension of the true data subspace is chosen correctly equal to three (for IPCR there is a marginal reduction in RMSET values when the dimension of the data subspace is taken

to be four, which may be attributed to numerical errors since error variances are being estimated in this method). It may be noted that in the IPCR method, 41 error variances are being estimated (assuming that the error covariance matrix along wavelengths is diagonal). The maximum number of error variances that can be estimated is equal to $(20 \times 21)/2 = 210$ for this data set. The results in Table 1 shows that the performance of proposed method IPCR is better than PCR, if the dimension of the data subspace is chosen to be three or more, even though both methods use the same information. MLPCR performs better than both PCR and IPCR, but it should be noted that this method requires knowledge of all the error variances, whereas IPCR estimates them from the same data set. A comparison of the standard deviations estimated using IPCR method with the true variances used in simulating data set 1, are shown in figure 3. From Fig. 3, it is observed that the estimated standard deviations are comparable to the true values and the Gaussian profile of the error variances is also estimated effectively.

The results of applying the three methods for data set 2 are shown in table 2. The minimum RMSET values for all three methods are obtained when the dimension of the data subspace is correctly taken to be three. Again it is observed that the performance of IPCR is better than PCR, but worse than MLPCR. For this data set, IPCR has been applied by assuming that the error variances vary only along the wavelength direction. It should be noted, however, that

Number of latent variables	Species	PCR	IPCR	MLPCR
1	A	0.208	0.198	0.159
	B	0.227	0.244	0.271
	C	0.203	0.190	0.182
	RMSET	0.213	0.212	0.210
2	A	0.138	0.185	0.077
	B	0.236	0.226	0.283
	C	0.152	0.163	0.086
	RMSET	0.181	0.193	0.176
3	A	0.042	0.026	0.018
	B	0.076	0.062	0.046
	C	0.047	0.025	0.017
	RMSET	0.057	0.041	0.030
4	A	0.045	0.028	0.018
	B	0.080	0.051	0.048
	C	0.049	0.024	0.017
	RMSET	0.060	0.036	0.031
5	A	0.049	0.032	0.018
	B	0.083	0.058	0.051
	C	0.051	0.025	0.019
	RMSET	0.063	0.041	0.033

Table 1: Comparison of different MVC models for simulated data set 1

for this data set the error variances vary with respect to wavelength as well as with respect to mixtures, which violates the assumptions of IPCR. Despite this, IPCR is able to provide better performance than PCR. The standard deviations estimated using IPCR are compared with the true standard deviations (averaged over all mixtures) in figure 4 when the dimension of the data subspace is correctly taken to be equal to three. The figure shows that the estimated standard deviations are comparable to the actual values, though the estimates are poorer in comparison to data set 1, due to violation of the assumption on the error covariance structure.

Table 3 contains results for data set 3 in which errors are correlated with each other. Although, Wentzell et al. (1997a) have developed a MLPCA method for correlated errors, we have applied the algorithm which is applicable for uncorrelated errors to investigate the effect of modelling assumptions on performance. The results show that all three methods have identical performance. This is due to the fact that the method that we have used for generating correlated errors tend to equalize the variances of errors in all measurements. The effect of correlation among different errors by itself does not appear to have an adverse impact on the performance of the methods.

4.3 Performance comparison on Experimental Data

We now present and discuss the results of applying PCR, IPCR, and MLPCR on the experimental data set 4 described in preceding section. This data set con-

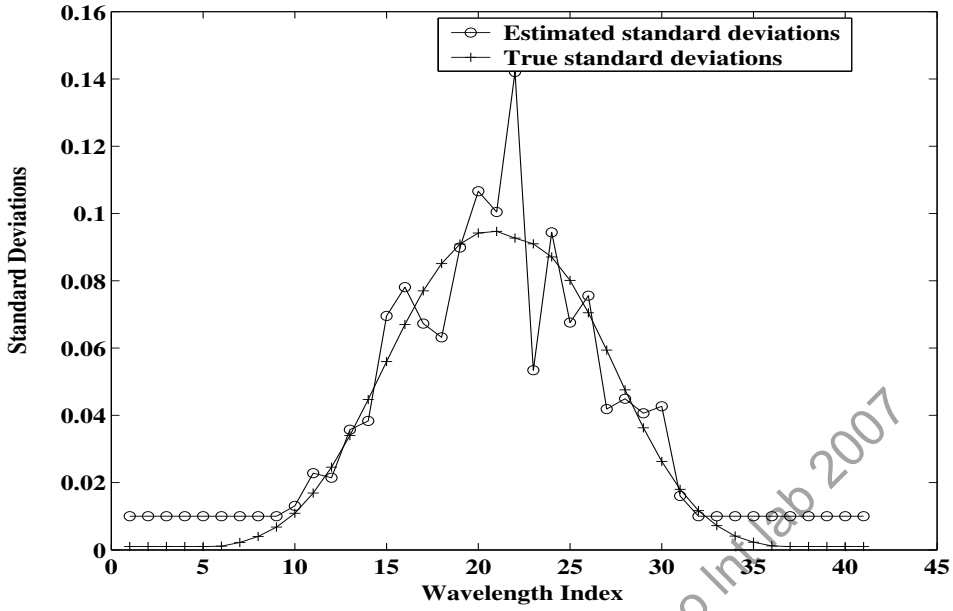

Figure 3: STD of errors at different wavelengths estimated using IPCA for data set 1

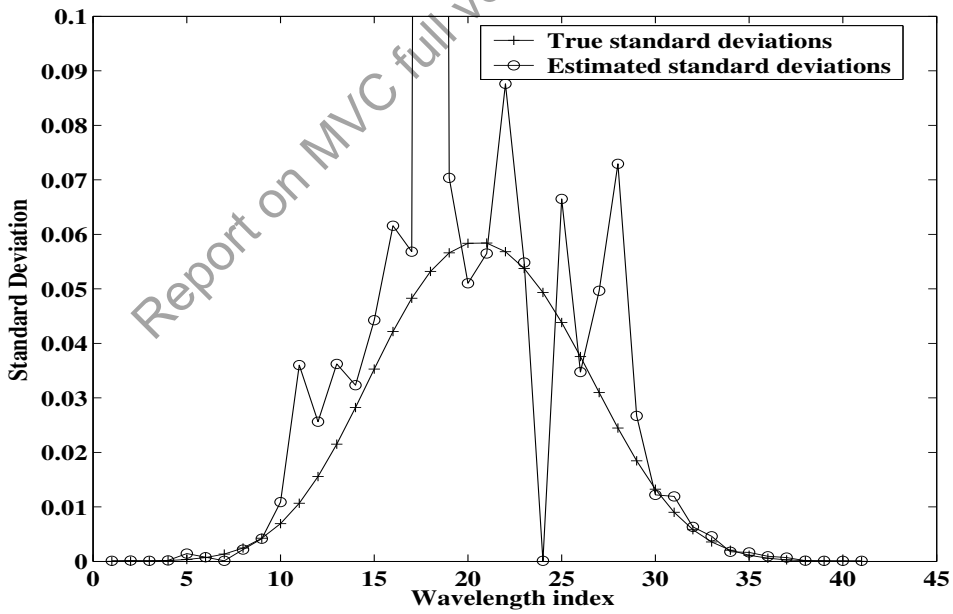

Figure 4: STD of errors at different wavelengths estimated using IPCA for data set 2

Number of latent variables	Species	PCR	IPCR	MLPCR
1	A	0.204	0.194	0.168
	B	0.227	0.230	0.297
	C	0.206	0.210	0.188
	RMSET	0.212	0.212	0.225
2	A	0.140	0.137	0.070
	B	0.237	0.234	0.295
	C	0.143	0.163	0.065
	RMSET	0.179	0.183	0.179
3	A	0.026	0.016	0.012
	B	0.039	0.031	0.022
	C	0.022	0.013	0.006
	RMSET	0.030	0.022	0.015
4	A	0.022	0.017	0.011
	B	0.039	0.031	0.023
	C	0.025	0.018	0.006
	RMSET	0.030	0.023	0.015
5	A	0.026	0.016	0.009
	B	0.041	0.029	0.024
	C	0.027	0.016	0.006
	RMSET	0.032	0.021	0.016

Table 2: Comparison of different MVC models for simulated data set 2

Number of latent variables	Species	PCR	IPCR	MLPCR
1	A	0.229		0.230
	B	0.186		0.186
	C	0.161		0.161
	RMSET	0.194		0.195
2	A	0.154	0.170	0.114
	B	0.194	0.198	0.184
	C	0.069	0.044	0.118
	RMSET	0.148	0.153	0.142
3	A	0.63×10^{-3}	0.001	0.001
	B	0.85×10^{-3}	0.001	0.001
	C	0.97×10^{-3}	0.001	0.001
	RMSET	0.83×10^{-3}	0.001	0.001
4	A	0.65×10^{-3}	0.001	0.65×10^{-3}
	B	0.91×10^{-3}	0.001	1.0×10^{-3}
	C	0.96×10^{-3}	0.001	0.65×10^{-3}
	RMSET	0.85×10^{-3}	0.001	0.88×10^{-3}

Table 3: Comparison of different MVC models for simulated data set 3

tains 26 distinct mixtures whose absorbance spectra have each been measured five times. Wentzell et al. (1997b) applied different MVC methods including PCR and MLPCR on this data set. In their approach, they treated all the 130 samples as distinct samples. This results in five different estimates of concentrations for each of the 26 mixtures. A more correct approach would be to first obtain the average of the five replicate spectra for each mixture and apply the MVC methods to the average spectra. As a first step, we apply PCR, and MLPCR, to the averaged spectra and compare it with the results obtained without averaging, to demonstrate the need for averaging when replicates are available. It should be noted that the size of the data matrix corresponding to

the use of averaged data is 26×176, and non-averaged case is 130×176. MLPCR is applied using the variances estimated directly from the five different replicates. The variance of the average spectra is computed by dividing this by the number of replicates. The results of applying the two methods are presented in Table 4. The results show that for MLPCR, the RMSET values do not change significantly beyond three factors (true subspace dimension), whereas for PCR the RMSET values decrease slowly after three factors. These results are consistent with those obtained by Wentzell et al. (1997b). The results also bring out a new and important feature. By comparing the RMSET values obtained for averaged data (26 spectra) with those for non-averaged data (130 spectra) we observe that a significant improvement in PCR performance is obtained by averaging. In the case of MLPCR, the improvement is marginal. The improvement occurs only when the number of latent factors is three or more. The reason for this can be attributed to the fact that by averaging, the error variances are reduced (by a factor equal to the number of replicates) and therefore, the adverse impact of heteroscedastic errors on PCA performance is reduced considerably. Thus when, replicates measurements are available, a simple approach of averaging can itself give significant improvement. The use of MLPCR method gives an additional improvement over that obtained due to averaging.

In this work, our focus is on situations when replicate measurements are not available. Therefore, all further comparisons are made on a data set consisting

Number of Latent factors	Species	PCR (130 samples)	PCR (26 samples average)	MLPCR (130 samples)	MLPCR (26 samples average)
1	Co	11.73	12.09	10.94	11.28
	Cr	3.50	3.60	3.06	3.15
	Ni	20.61	21.21	24.26	25.00
	Total	13.84	14.25	15.46	15.94
2	Co	8.45	8.82	7.14	7.53
	Cr	3.16	3.33	3.05	3.22
	Ni	11.75	12.29	16.90	17.86
	Total	8.56	8.95	10.74	11.34
3	Co	8.32	0.77	0.48	0.28
	Cr	3.17	0.40	0.13	0.09
	Ni	11.84	1.55	0.63	0.34
	Total	8.55	1.03	0.46	0.26
4	Co	8.35	0.79	0.48	0.28
	Cr	3.19	0.28	0.13	0.09
	Ni	11.91	1.24	0.61	0.33
	Total	8.60	0.87	0.45	0.26
5	Co	5.55	0.83	0.26	0.28
	Cr	2.35	0.27	0.09	0.09
	Ni	7.75	1.29	0.44	0.35
	Total	5.67	0.90	0.30	0.27
6	Co	2.35	0.87	0.23	0.14
	Cr	0.75	0.27	0.08	0.06
	Ni	3.12	1.26	0.43	0.29
	Total	2.30	0.90	0.28	0.19

Table 4: Comparison of PCR and MLPCR on averaged and non-averaged data for experimental data set

of non-replicated measurements of spectra of 26 mixtures. This data set is constructed by randomly picking one mixture spectra from each of the five replicates of the original data set.

Table 5 shows the results of applying PCR, MLPCR and IPCR on two different random sets of 26 mixtures drawn from the experimental data set. IPCR is applied under the assumption that the error variances vary with respect to mixtures, but are constant with respect to wavelengths. Thus, 26 different variances have to be estimated in this method. The results of Table 5 show that the performance of PCR and IPCR are comparable, while the MVC model developed using MLPCR has much better predictive capability (if the number of latent factors are chosen to be three or more). The utilization of error variances has resulted in better performance of MLPCR. However, although IPCR attempts to take into account differences in error variances across mixtures by estimating these along with the subspace, no improvement in performance over PCR is obtained. This could be due to the fact that the predominant variation of error variation for this data set is along wavelength direction and not along mixture direction.

In order to verify the above hypothesis, we apply IPCR to the same two random data sets, by assuming that the error variances vary along wavelength direction, but are constant along mixture direction. In this case, it is necessary to estimate 176 error variances simultaneously along with the data subspace in the first step of MVC model development, by using the transpose of the 26×176

Number of Latent factors	Species	PCR Set1	PCR Set2	IPCR Set1	IPCR Set2	MLPCR Set1	MLPCR Set2
1	Co	12.30	12.01	12.30	12.03	11.43	11.22
	Cr	3.58	3.68	3.59	3.68	3.12	3.18
	Ni	21.15	21.06	21.14	21.02	24.95	24.95
	Total	14.28	14.16	14.27	14.14	15.95	15.90
2	Co	9.45	8.84	9.51	6.93	7.67	7.46
	Cr	3.30	3.49	3.30	3.83	3.19	3.26
	Ni	13.97	13.77	14.49	18.12	17.96	17.70
	Total	9.92	9.66	10.19	11.42	11.43	11.25
3	Co	10.07	8.60	8.83	7.04	0.66	0.70
	Cr	3.27	3.61	2.71	3.04	0.19	0.15
	Ni	13.56	14.08	12.70	15.12	0.70	1.06
	Total	9.93	9.75	9.07	9.79	0.56	0.74
4	Co	8.75	8.65	6.93	7.95	0.68	0.70
	Cr	3.27	3.61	2.71	3.04	0.19	0.15
	Ni	13.56	14.08	12.70	15.12	0.70	1.06
	Total	9.93	9.75	9.07	9.79	0.56	0.74
5	Co	6.22	6.86	7.36	5.34	0.32	0.68
	Cr	2.46	1.84	2.75	3.20	0.10	0.15
	Ni	9.86	6.28	13.32	15.20	0.45	1.13
	Total	6.88	5.48	8.93	9.48	0.33	0.77
6	Co	5.56	3.24			0.67	0.68
	Cr	1.89	1.28			0.19	0.15
	Ni	6.73	4.33			0.67	1.10
	Total	5.16	3.21			0.56	0.75

Table 5: Comparison of MVC models on two random subsets of experimental data

data matrix. It was found that the optimization algorithm did not converge due to the large number of optimization variables. A sub-optimal approach we have developed to resolve this problem is as follows. We first divide the 176×26 data matrix into sub-matrices of dimensions not exceeding 50 rows each. By treating each sub-matrix separately, we are required to estimate only 50 error variances simultaneously. Since, the subspace estimated for each sub-matrix will not be identical, we need to estimate a unique subspace. For this purpose, once we estimate all of the error variances (corresponding to different wavelengths) by above sub-optimal approach, we collate all of them to construct the error covariance matrix and use it to transform the entire 176×50 data matrix as in Eq. 10. A singular value decomposition of this transformed data can be used to estimate the true data subspace as described in subsection 2.3. It should be noted that this last step is similar to applying MLPCA using known error covariance matrix, with the difference that this error covariance matrix is estimated by applying the IPCR method to each sub-matrix separately. The sub-optimal IPCR approach described here is a general strategy that can be used for all data sets where it s required to estimate a large number of error variances. The results of applying this method to the random data sets drawn from the experimental data set are presented in Table 6. The results clearly show that the MVC model developed using IPCR has as good predictive capabilities as the MLPCR model (compare with last two columns of Table 5), if we choose the number of factors

Number of latent factors	Species	IPCR Set 1	IPCR Set 2
3	Co	0.68	0.76
	Cr	0.16	0.13
	Ni	1.67	2.59
	Total	1.04	1.56
4	Co	0.13	0.18
	Cr	0.09	0.11
	Ni	0.47	0.51
	Total	0.29	0.32
5	Co	0.14	0.19
	Cr	0.08	0.12
	Ni	0.46	0.46
	Total	0.28	0.30
6	Co	0.17	0.17
	Cr	0.09	0.10
	Ni	0.55	0.39
	Total	0.34	0.24

Table 6: Performance of IPCR on experimental data assuming error std varies with wavelength

equal to four or more. It can be noted that the best performance is obtained if we choose the number of factors to be one more than the actual number of species in the mixture. The extra factor could be due to the differences in error variances along sample direction (which has not been accounted for in the model) being artificially modelled as an additional factor.

The estimated values of standard deviations using the above sub-optimal approach when the number of latent factors are chosen to be three, is shown in figure **??**.

As shown in the graph, the estimated error standard deviations using the IPCR method follows the boat shaped trend in the 'true' standard deviations es-

timated from the replicate measurements. These results show that the proposed method is able to perform almost as well as the MLPCR method without the need for replicate measurements. The results also confirm that the differences in error variances in the wavelength direction are more dominant than those along the mixture direction.

5 Summary

We have proposed a new approach for developing multivariate calibration models from non-replicated measurements when the error variances in absorbance measurements vary along one mode. In particular, we note that the dominant mode of variation of error variances is usually along wavelength direction. The MVC model developed using the proposed approach has better prediction accuracies than that obtained using PCR, and approaches the accuracy of the MVC model developed using maximum likelihood PCR method which requires replicate measurements. For applying the proposed method to data sets which require a large number of error variances to be estimated, an approach has been developed, based on dividing the data set into manageable subsets.

References

[1] Wentzell, P. D., Andrews, D. T., and Kowalski, B. R. (1997b) Maximum likelihood multivariate calibration. *Analytical Chemistry*, **69**, 2299–2311.

[2] Rao, C. R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya, Ser. A*, **26**, 329–358.

[3] Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K., and Kowalski, B. R. (1997) Maximum likelihood principal component analysis. *J. Chemometrics*, **11**, 339–366.

[4] Narasimhan, S. and Shah, S. L. (2003) Model identification and error covariance matrix estimation from noisy data using pca. *International Symposium on Advanced Control of Chemical Processes, ADCHEM 2003, Hong Kong*.

[5] Paatero, P. and Tapper, U. (1993) Analysis of different modes of factor analysis as least squares fit problems. *Chemometrics Intell. Lab. Syst.*, **18**, 183–194.

[6] Fuller, W. (1987) *Measurement Error Models*. Wiley, New York.

[7] Chan, N. N. and Mak, T. K. (1983) Estimation of multivariate linear functional relationships. *Biometrika*, **70**, 263–267.

[8] Cheng, C. L. and Van Ness, J. W. (1994) On estimating linear relationships when both variables are subject to errors. *J.R. Statist. Soc.*, **56**, 167–183.