Multivariate calibration of non-replicated measurements for the factored noise model

Nirav Bhatt[1], Shankar Narasimhan*

Department of Chemical Engineering, Indian Institute of Technology-Madras, Chennai, India 600 036

Abstract

The accuracy of a multivariate calibration (MVC) model for relating concentrations of multicomponent mixtures to their spectral measurements, depends on effective handling of errors in the measured data. For the case when error variances vary along only one mode (either along mixtures or along wavelengths), a method to estimate the error variances simultaneously along with the spectral subspace was developed by Narasimhan and Shah (*Control Engineering Practice*, **16**, (2008), 146-155). This method was exploited by Bhatt *et al.* (*Chemom. Intell. Lab. Syst.*, **85**, (2007), 70-81) to develop an iterative principal component regression (IPCR) MVC model, which was shown to be more accurate than models developed using PCR. In this work, the IPCR method is extended to deal with measurement errors whose variances vary along both modes, by using a factored noise model. As a first step, an iterative procedure is developed to estimate the error variance factors along with the spectral subspace, which is subsequently used in developing the regression model. Using simulated and experimental data, it is shown that the quality of the MVC model developed using the proposed method is better than that obtained using PCR, and is as good as the model obtained using Maximum Likelihood PCR, which requires knowledge of the error variances. For dealing with large data sets, a sub-optimal approach is also proposed for estimating the large number of error variances.

*Corresponding author

Email address: naras@iitm.ac.in (Shankar Narasimhan)

[1]Currently a doctoral student at École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Key words:
Multivariate calibration; Principal component regression; Error covariance matrix; Factorial noise model; Heteroscedastic errors.

1. Introduction

Multivariate calibration (MVC) methods are routinely used in analytical chemistry for the development of a predictive model relating properties of chemical mixtures such as concentrations to its spectroscopic measurements. The MVC methods include multiple linear regression (MLR), principal component regression (PCR), partial least-squares regression (PLS) etc. Among these, PCR is one of the widely used methods to develop a MVC model. PCR develops a MVC model using a two-step process. In the first step, principal component analysis (PCA) is used to estimate a lower dimension subspace from the spectroscopic data. The measured spectra are projected on to this estimated subspace, and the weights (scores) for representing the projections in terms of the basis for the estimated subspace are obtained. In the second step, a multivariate linear regression model is developed between the concentrations of the calibration mixtures and their scores. The developed MVC model can be used to predict the concentration of species for a new mixture from its measured absorbance spectra.

If the error variances change along only one mode, that is, the error variances change with respect to mixtures or with respect to wavelengths, then it is known that the optimal spectral subspace can be obtained by first transforming the measured data using the inverse of the cholesky factor of the error covariance matrix and applying PCA to these transformed measurements [3]. It may be noted, that if the errors are uncorrelated, then the error covariance matrix is diagonal and this transformation is identical to a specific scaling of the measured data. Cochran and Horne [4] developed a statistically weighted principal component analysis (SWPCA) method using a scaling technique, when the error variances vary with respect to samples as well as wavelengths. A simplified model of the error variances is assumed in this approach by expressing the variance of measurement error of a sample i at a specific wavelength j as a product of two factors. The first factor is the contribution from mixture i which is same for the absorbances at all wavelengths in this mixture, and the second factor is the contribution from wavelength j which is same for the absorbance measurements at this wavelength in all mixtures. This type of model for error variances is denoted as

the factored noise model. In the SWPCA method, both the rows and the columns of the noisy data matrix are scaled using *a priori* diagonal scaling matrices, whose elements are the square root of the factors of the noise variances, before applying PCA. A further interesting feature of these methods applied to scaled/transformed data is that they give rise to a specific pattern of eigenvalues, which can be used to estimate the true rank of data matrix and the true subspace.

For more general noise models, Wentzell *et al.* [5] developed a method called maximum likelihood PCA (MLPCA), which incorporates information of measurement error variances to obtain a better estimate of the true subspace. MLPCA can be viewed as an iterative method which transforms the subspace identified by PCA iteratively to obtain the maximum likelihood estimates of variables. In contrast to SWPCA, MLPCA can deal with errors which have different variances (heteroscedastic errors) and which may be correlated. Using MLPCA as the first step of PCR, Wentzell *et al.* [6] developed a maximum likelihood PCR (MLPCR) calibration model, which has better prediction accuracy as compared to the PCR model.

SWPCA and MLPCA (and consequently MLPCR) require complete information about the error variances to be specified *a priori*. Such information can be obtained by performing replicate experiments for all measurements and estimating the error variances from the sample variances. In many situations, replicate measurements may cost significant resources and time. If the measurement error variances and covariances can be estimated from non-replicated noisy data along with the true subspace, the cost and time required for performing replicate measurements can be saved.

Narasimhan and Shah [1] developed a new method called iterative principal component analysis (IPCA) which can estimate simultaneously the error covariance matrix and the true subspace from non-replicated measurements for a particular noise model. In this method, it is assumed that the error variances can vary along only one mode. IPCA employs an iterative procedure combining PCA with maximum likelihood estimation (MLE) estimation of the error variances. Similar to SWPCA, IPCA uses the error variances to scale the measurements prior to applying PCA to estimate the true data subspace. It can be regarded as a special case of SWPCA, in which the error variance factors along one mode are set equal to unity and the factors along the other mode are estimated. The dimension of the true data subspace is estimated by examining the singular values of the scaled data matrix. Recently IPCA has been used to develop a MVC model for non-replicated

measurements, which is known as iterative principal component regression (IPCR) [2]. It was demonstrated that MVC models built using IPCR has better prediction accuracy than PCR models.

Although, IPCR is a useful approach for developing accurate MVC models when replicate measurements are not available, the noise model assumed in IPCA is restrictive, since it assumes that error variances can vary along only one mode. It would be worthwhile to generalize this approach to the case of the factored noise model, which allows the error variances to vary along both modes. In this case, it is necessary to estimate the error variance factors of the factorial model simultaneously along with the true subspace from non-replicated measurements. In this paper, a new approach referred to as Iterative Weighted Principal Component Analysis (IWPCA) is proposed for this purpose. Using IWPCA as the first step, a method known as iterative Weighted Principal Component Regression (IWPCR) is also proposed for developing a MVC model and its prediction accuracy is compared with other MVC approaches.

Section 2 describes the theoretical development of IWPCA and IWPCR. The predictive ability of the developed MVC model using IWPCR is evaluated on simulated as well as experimental data sets. In Section 3, a description of simulated data sets and experimental data sets is provided. The quality of IWPCR calibration model is compared with PCR and MLPCR in Section 4. The main contributions of this paper are summarized in Section 5.

2. Theory

In the development of a MVC model using PCR, the first step is to estimate the true data subspace from the noisy measurements. Principal Components Analysis (PCA) is a widely used technique for this purpose. There are several variants of PCA depending on the assumptions made regarding the errors in the measurements. Table 2 gives a summary of the various methods and assumed structure of measurement error variances and their availability. The method IWPCA listed in the last row of this table is the one proposed in this paper.

The IWPCA method proposed in this paper is an approach for estimating the true data subspace from noisy measurements, while simultaneously estimating error variances for a factorial noise model. The development of this method combines ideas from the SWPCA method developed by Cochran

Table 1: Summary of different PCA variants corresponding to different measurement error variance models

Sr. No.	Name of method	Error variance structure	Comment
1	PCA	$\sigma_{ij}^2=\sigma^2$	Measurement errors are independent and identically distributed
2	SWPCA	$\sigma_{ij}^2=x_i z_j$	Each error variance is a product of two factors x_i and z_j x_i=factor accounts for variable direction z_j=factor accounts for observation direction
3	MLPCA	σ_{ij}^2	Errors may have different variances and maybe uncorrelated or correlated. Error variances and covariances are available *a priori*
4	IPCA	$\sigma_{ij}^2=\sigma_i^2$ or σ_j^2	Errors vary along either variable direction or observation direction. Error variances are estimated from the data
5	IWPCA	$\sigma_{ij}^2=x_i z_j$	Error variances vary along variable and observation directions Error variances are estimated from the data

*σ_{ij}^2= variance of error in measurement of variable i and observation j,

and Horne [4] and the IPCA method developed by Narasimhan and Shah (2007).

2.1. Problem formulation

Let $\mathbf{y}^*(j) : m \times 1$ represent the true values of m variables at sampling instant j, which are linearly related by p independent equations:

$$\mathbf{A}\mathbf{y}^*(j) = 0 \tag{1}$$

where $\mathbf{A}$: $(p \times m)$ is a constraint matrix. The rows of $\mathbf{A}$ form a basis for a p-dimensional subspace of $\mathbf{R}^m$, while the true data vectors span a $(m-p)$-dimensional subspace of $\mathbf{R}^m$, orthogonal to the row space of $\mathbf{A}$. Measurements $\mathbf{y}(j)$ of all the variables corrupted by the random noises (errors) are available at each sampling instant j, and can be written as:

$$\mathbf{y}(j) = \mathbf{y}^*(j) + \epsilon(j) \tag{2}$$

For n such measurements, $\mathbf{y}(1), \mathbf{y}(2)...., \mathbf{y}(n)$, an $(m \times n)$ data matrix $\mathbf{Y}$ can be defined as

$$\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(n)] = \mathbf{Y}^* + \boldsymbol{\epsilon} \tag{3}$$

In the case of spectroscopic data, element $y_i(j)$ is the absorbance of a mixture i at wavelength j, and element $\epsilon_i(j)$ is the random error in the absorbance measurement $y_i(j)$. Generally, the number of mixtures used in developing the calibration model is much less than the number of wavelengths at which the absorbances are measured. Thus the rank of the measured data matrix is m.

In general, the errors in different measurements may have different variances and may also be correlated. In this paper, we consider the factored noise model for the random errors. This noise model assumes that the random errors have an expected value of zero and a variance given by.

$$var(\epsilon_i(j)) = \sigma_{ij}^2 = x_i z_j \tag{4}$$

where x_i and z_j are factors corresponding to contributions to the error variance from mixture i and wavelength j, respectively. The random errors are also assumed to be mutually independent and are also independent of the true values. The factored noise model allows the error variances to vary with respect to both mixtures and wavelengths, but imposes a structure by parameterizing them in terms of $m + n$ factors.

The objective in this work is to simultaneously estimate the $(m-p)$-dimensional subspace of $\mathbf{R}^m$ in which the true data vectors lie, and the error variance factors from the measured data. In order to motivate our development, we first consider the case when the error variance factors are known and it is required to estimate a basis for the true data subspace. For this purpose the statistically weighted PCA method developed by Cochran and Horne [4] can be used, which is described in the following subsection.

2.2. Statistically weighted principal component analysis

The Statistically Weighted PCA (SWPCA) method developed in [4] applies PCA to appropriately weighted measurements in order to estimate the true data subspace. This method provides a consistent estimate of the true data subspace if the random errors in measurements follow the factored noise model. The method is described briefly in this section.

Let two non-singular diagonal matrices, $\mathbf{B}$ and $\mathbf{D}$ be defined as:

$$\begin{aligned} \mathbf{B} &= diag(b_1, b_2, .., b_m) \\ \mathbf{D} &= diag(d_1, d_2, .., d_n) \end{aligned} \tag{5}$$

Using the above matrices $\mathbf{B}$ and $\mathbf{D}$, the weighted measurement matrix $\mathbf{Y}_w$ is defined as:

$$\mathbf{Y}_w = \mathbf{BYD} \tag{6}$$

If the measurement error variances follow the factored noise model defined by 4, then by choosing the scaling matrices $\mathbf{B} = (\mathbf{X})^{-1/2}$ and $\mathbf{D} = (\mathbf{Z})^{-1/2}$, Cochran and Horne [4] showed that the expected value of the covariance matrix $\mathbf{Y}_w$ ($\mathbf{M}_w$) is given by

$$E(\mathbf{M}_w) = \left(\frac{1}{n}\right) \mathbf{X}^{-1/2}\mathbf{Y}^*\mathbf{Z}^{-1}\mathbf{Y}^*\mathbf{X}^{-1/2} + \mathbf{I} \tag{7}$$

It can be easily shown that $E(\mathbf{M}_w)$ and the first term on the right hand side have the same eigenvectors. Since the true data lie in a $m-p$ dimensional subspace of $\mathbf{R}^m$, the matrix in the first term of Eq. (7) has p zero eigenvalues, while the identity matrix in the second term has all eigenvalues equal to unity. Thus, the least p eigenvalues of $E(\mathbf{M}_w)$ should be equal to 1. This implies, that by examining the pattern of eigenvalues of $\mathbf{M}_w$, the rank of true data matrix (the dimension of subspace in which the true data lies) can be determined. A basis for the true data subspace can be

obtained from the eigenvectors corresponding to the eigenvalues which are greater than unity. Cochran and Horne [4] assume that the factors of the error variances are known. Since such information regarding error variances are difficult to obtain or require repeated measurements to be made, it would be advantageous if these factors can be estimated without the need for replicate measurements. In the next section, a new method is proposed, which simultaneously estimates the factors x_i and z_j and the true data space.

2.3. Iterative weighted principal component analysis

Given measurements of a process which can be described by Eqs. (1 and 2), Narasimhan and Shah [1] developed a method called iterative PCA (IPCA) which can simultaneously estimate elements of error covariance matrix and the true data subspace. This method assumes that the error variances can vary only along either the sample direction or variables direction. Using, the estimated error variances, the measurements are scaled before applying PCA to estimate the true data subspace. The noise model assumed in IPCA is restrictive. A more general noise model should allow for different error variances for different measurements. However, this leads to a large number of variances ($m \times n$ even for the case of uncorrelated errors) to be estimated from the data, which may not be possible due to limitations in computing power as well as on the total number of variances that can be estimated. A compromise solution is to make use of the factored noise model of Cochran and Horne [4]. Using this model, only $(m+n)$ variances need to be estimated, which may be possible from non-replicated measurements. In this section, the mathematical details of the proposed method called iterative weighted PCA (IWPCA) are described. The proposed method combines concepts and ideas drawn from SWPCA and IPCA for simultaneously estimating the true data subspace and error variance factors of the factored noise model given by Eq. (4). Without loss of generality, instead of estimating a basis for the true data subspace, we will estimate a basis for the row space of the constraint matrix $\mathbf{A}$, which is orthogonal to the true data subspace.

We start with the assumption that the dimension p for the row space of constraints is known. An initial estimate of $\mathbf{A}$, say $\hat{\mathbf{A}}^0$ is assumed to be available. Such an initial estimate can be obtained by applying PCA to the measurements and using the eigenvectors corresponding to the smallest p eigenvalues. The constraint residuals at each instant j are computed using the initial estimate $\hat{\mathbf{A}}^0$ as

$$\mathbf{r}(j) = \hat{\mathbf{A}}^0\mathbf{y}(j) = \hat{\mathbf{A}}^0\mathbf{y}^*(j) + \hat{\mathbf{A}}^0\epsilon(j) \tag{8}$$

In the next step, we attempt to estimate the factors of the error variances from the constraint residuals. If the estimated model $\hat{\mathbf{A}}^0$ for the row space of constraints is exact, then the weighted residuals $\mathbf{r}(j)/z_j$ will be independently and identically distributed multivariate normal variables given by

$$\mathbf{r}(j)/z(j) \sim \mathcal{N}(0, \hat{\mathbf{A}}^0 \mathbf{\Sigma} (\hat{\mathbf{A}}^0)^T) \tag{9}$$

where $\mathbf{\Sigma}$ is a $m \times m$ diagonal matrix whose diagonal elements are given by $(x_1, x_2, \ldots x_n)$. Since the errors are independent, the likelihood function of $\mathbf{r}(1)/z(1), \mathbf{r}(2)/z(2), ...\mathbf{r}(n)/z(n)$ can be constructed which can be maximized to obtain an estimate of the factors x_i and z_j. This leads to the following optimization function.

$$\min_{x_i, z_j} n \log|\hat{\mathbf{A}}^0 \mathbf{\Sigma} (\hat{\mathbf{A}}^0)^T| + \sum_{j=1}^{n} (\mathbf{r}(j)/z(j))^T (\hat{\mathbf{A}}^0 \mathbf{\Sigma} (\hat{\mathbf{A}}^0)^T)^{-1} (\mathbf{r}(j)/z(j)) \tag{10}$$

The above non-linear optimization function is used to obtain the estimates of x_i and z_j factors. Positivity constraints can also be imposed on the estimated factors to ensure that the variances are positive. The number of variance factors that can be estimated using Eq. (10) depends on the rank of the sample covariance matrix of constraints residuals. For the case when the noise variances vary along only one mode, Narasimhan and Shah (2008) showed that the maximum number of elements of the error covariance matrix that can be estimated is $p(p+1)/2$ where p is the rank of the constraint residuals covariance matrix. The limitation arises from the fact that we are estimating the error variances using the $p(p+1)/2$ variances and covariances of the symmetric constraint residual covariance matrix. The same argument can be used using the sample covariance of the weighted constraint residuals even for the factored noise model. Thus, we can estimate the parameters of the factored noise model if $m + n$ is less than or equal to $p(p+1)/2$.

The estimated error variance factors can be used for obtaining the weighted data matrix, $\mathbf{Y}_w$, defined by Eq. (6), by assigning $\mathbf{B} = (\mathbf{X})^{-1/2}$ and $\mathbf{D} = (\mathbf{Z})^{-1/2}$. PCA is applied to the weighted data matrix to get a new estimate of the constraint matrix. Let the notation $[\mathbf{U}_w, \mathbf{S}_w, \mathbf{V}_w] = svd(\mathbf{Y}_w, m)$ be used to denote the truncated singular value decomposition of $\mathbf{Y}_w$, where $\mathbf{U}_w$ is an $m \times m$ matrix of left singular vectors, $\mathbf{S}_w$ is an $m \times m$ diagonal matrix containing m non-zero singular values ordered from the largest to the smallest, and $\mathbf{V}_w$ is an $m \times n$ matrix of right singular vectors. Using this, the

matrix $\mathbf{Y}_w$ can be written as

$$\mathbf{Y}_w = \mathbf{U}_{1w}\mathbf{S}_{1w}\mathbf{V}_{1w}^T + \mathbf{U}_{2w}\mathbf{S}_{2w}\mathbf{V}_{2w}^T \tag{11}$$

where $\mathbf{U}_{1w}$, $\mathbf{S}_{1w}$, and $\mathbf{V}_{1w}$ correspond to the first $(m-p)$ largest singular values, while $\mathbf{U}_{2w}$, $\mathbf{S}_{2w}$, and $\mathbf{V}_{2w}$ correspond to the smallest p singular values of the $\mathbf{Y}_w$. The columns of $\mathbf{U}_{1w}$ and $\mathbf{V}_{1w}$ form a basis for the $(m-p)$ dimensional estimated mixture and estimated spectral subspaces in the transformed domain, respectively. The new estimated constraint matrix in the original space is obtained as

$$\hat{\mathbf{A}} = \hat{\mathbf{A}}_w\mathbf{B} \tag{12}$$

The entire procedure is repeated until estimates of the constraint matrix and variance factors converge. If the average of the last p eigenvalues calculated using PCA does not change significantly from one iteration to the next, then the iterative procedure is deemed to have converged. Following the arguments made in SWPCA, if the estimated variance factors after convergence of IWPCA are close to their true values, the last p eigenvalues of covariance matrix should be equal to unity.

The iterative procedure for estimating x_i and z_j and the constraint matrix ($\mathbf{A}$) can be summarized as follows.

Step 1: Initialize iteration (k=1) and sum of eigenvalues $\lambda^k = 0$.

Step 2: Set initial guess $\hat{\mathbf{A}}^0$ to be the constraint matrix estimated using PCA on the data matrix $\mathbf{Y}$.

Step 3: Obtain the solution for $\hat{x}_i^k$ and $\hat{z}_j^k$ by minimizing the non-linear optimization function in Eq. (10)

Step 4: Set $\hat{\mathbf{X}}^k = \text{diag}(\hat{x}_1^k, \hat{x}_2^k, ..., \hat{x}_m^k)$ and $\hat{\mathbf{Z}}^k = \text{diag}(\hat{z}_1^k, \hat{z}_2^k, ..., \hat{z}_n^k)$ and define the weighting matrices $\hat{\mathbf{B}}^k = \hat{\mathbf{X}}^{k(-1/2)}$ and $\hat{\mathbf{D}}^k = \hat{\mathbf{Z}}^{k(-1/2)}$

Step 5: Obtain the truncated svd of weighted data matrix $\mathbf{Y}_w^k = \mathbf{B}^k\mathbf{Y}\mathbf{D}^k$ as:

$$[\mathbf{U}_w^k, \mathbf{S}_w^k, \mathbf{V}_w^k] = svd(\mathbf{Y}_w^k, m)$$

Estimate the constraint matrix, $\hat{\mathbf{A}}^k = (\mathbf{U}_{2w}^k)^T\hat{\mathbf{B}}$, where $\mathbf{U}_{2w}^k$ is the sub-matrix of $\mathbf{U}_w^k$ corresponding to the last p columns.

Step 6: Let λ^k be the sum of the last p singular values. If relative change in λ is less than specified tolerance then stop else continue.

Step 7: Increment iteration counter k and use the estimated values of x_i and z_j factors in the previous iteration as updated guesses for the next iteration and return to step 3.

It should be noted that although the above description assumes that all the factors for the error variances are estimated simultaneously, the same procedure can be applied even if some of the factors (say the x_i factors or the z_j factors) are known *a priori* and only the remaining have to be estimated. In such cases, the optimization in step 3 is carried out only for the unknown factors. It may be noted that when the factors along one of the modes are assumed to be known or kept fixed at some known values during the optimization in Step 3, then the above procedure is identical to the application of the IPCA algorithm as described in Narasimhan and Shah [1] and Bhatt et al. [2].

2.3.1. Handling of large data sets

In spectroscopic data, the absorbances of mixtures are usually measured over a large number of wavelengths, and this can require the estimation of a large number of variance factors. It may sometimes be difficult to estimate such a large number of variables using the nonlinear optimization approach in Step 3. In our experience with different data sets, we have found that it is possible to estimate about 50 error variances simultaneously. For larger data sets where it is required to estimate more error variances, a two-step sub-optimal approach is proposed to estimate the variance factors in Step 3. In the first step, it is assumed that error variances vary along one mode (generally, along the sample direction) and the above algorithm is applied to the data matrix to estimate these error variances. The estimated error variances are designated as one of the factors (say x_i). In the second step, the factors along the other mode (wavelength direction) (z_j) are estimated by keeping the factors x_i fixed. If the optimization along the second mode is still large dimensional, then it can be further broken down into smaller optimization problems by sub-dividing the data matrix into the several sub-matrices along the mode whose variance factors are being estimated, and estimating the subset of variance factors using the same procedure as described above. Once all the variance factors are estimated, they can be collated and the SWPCA method used to estimate a unique true lower dimensional subspace. Although, this approach is sub-optimal it allows us to handle any large data matrix that may arise in practical applications.

2.4. Development of Multivariate calibration model

After an estimate of the true data subspace is obtained using IWPCA, the measured absorbance spectra have to be projected on to this subspace. Unlike PCA, since the variances of errors in different absorbances are not identical, orthogonal projections are not optimal. The maximum likelihood estimate of the absorbance matrix is given by

$$\hat{\mathbf{Y}}_{MLE} = \hat{\mathbf{Z}}^{0.5}\hat{\mathbf{U}}_{1w}\hat{\mathbf{S}}_{1w}\hat{\mathbf{V}}_{1w}^{T}\hat{\mathbf{X}}^{0.5} \tag{13}$$

where $\hat{\mathbf{Z}}$ and $\hat{\mathbf{X}}$ are diagonal matrices containing the estimated variance factors along the mixture and spectral directions, respectively, while the matrix $\hat{\mathbf{U}}_{1w}\hat{\mathbf{S}}_{1w}\hat{\mathbf{V}}_{1w}^{T}$ is the first term as defined in 11 obtained at convergence of IWPCA.

The regression matrix relating the estimated absorbance spectra to the mixture concentrations can be obtained by first determining an orthogonal basis for the absorbance subspace and the corresponding scores for representing the estimated spectra in terms of this basis using the svd of $\hat{\mathbf{Y}}_{MLE}$ as follows.

$$svd(\hat{\mathbf{Y}}_{MLE}, s) = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{T} = \hat{\mathbf{T}}_{MLE}\hat{\mathbf{V}}^{T} \tag{14}$$

where $s = m - p$ is the number of species in the mixtures (rank of $\hat{\mathbf{Y}}_{MLE}$), $\hat{\mathbf{T}}_{MLE}$ is the $m \times s$ scores matrix and $\hat{\mathbf{V}}$ is the $n \times s$ orthonormal basis for the estimated absorbance subspace.

The linear regression model between concentrations and the scores is given by

$$\mathbf{C} = \hat{\mathbf{T}}_{MLE}\boldsymbol{\beta} + \mathbf{F} \tag{15}$$

where $\mathbf{C}$ is an $m \times s$ measured concentration matrix which consists of s species in the m calibration mixtures and $\boldsymbol{\beta}$ is the regression matrix and $\mathbf{F}$ is an $m \times s$ matrix of errors. The least squares solution of Eq. (15) for the regression matrix is given by

$$\boldsymbol{\beta} = (\hat{\mathbf{T}}_{MLE}^{T}\hat{\mathbf{T}}_{MLE})^{-1}\hat{\mathbf{T}}_{MLE}^{T}\mathbf{C} \tag{16}$$

Since the above approach uses the error variance factors as well as the mixture/absorbance subspace estimated using IWPCA, the MVC approach is denoted as IWPCR to maintain consistency with other PCA based calibration methods.

Using the regression model, the concentrations of a new mixture can be predicted from its measured absorbance spectrum. Let $y_{new} : 1 \times n$ be

the measured absorbance spectrum of a new mixture. The scores of a new spectrum are obtained using the maximum likelihood projection as

$$t_{new} = y_{new}\hat{\mathbf{Z}}^{-1}\hat{\mathbf{V}}(\hat{\mathbf{V}}^T\hat{\mathbf{Z}}^{-1}\hat{\mathbf{V}})^{-1} \tag{17}$$

and the unknown concentrations in the new mixture are predicted as

$$c_{new} = t_{new}\boldsymbol{\beta} \tag{18}$$

3. Data sets used in evaluating MVC models

The MVC models developed using IWPCR method is evaluated using three simulated data sets and two experimental data sets consisting of mixtures of three species. A comparative evaluation of IWPCR method is performed by applying PCR and MLPCR methods to the same data sets. Due to space restriction, the PCR and MLPCR methods are not described here and the reader is referred to the papers by [6] and [2] for the details of these methods.

3.1. Simulated data sets

Simulated data sets for the mixture of three species are generated following a similar procedure as described in [6]. First, the noise free data sets are generated as follows:

1. The spectral profiles of the three species are taken to be Gaussian distributions with a peak absorbance at 460 nm, 500 nm, and 540 nm, respectively, and a standard deviation of 20 nm. Pure component spectral vectors are generated between 400 nm and 600 nm at intervals of 5 nm to obtain a 3×41 pure component spectra matrix.
2. Concentrations of the above three species for 20 mixtures are generated by choosing random numbers between 0 and 1 from a uniform distribution for each species. In this manner, a 20×3 concentration matrix is obtained
3. The noise free data (true data) absorbance matrix of dimension 20×40 is obtained by multiplying the concentration matrix and pure component spectra matrix.

The noisy data sets are generated by adding random noise (errors) to the true absorbance data. Three different error covariance structures are used to

obtain three different data sets. In data set 1, the error variances in different absorbance measurements are assumed to follow the factorial noise model. The variance factors in the mixture direction (x_i) are assigned as 5% of the maximum absorbances for the corresponding mixtures. The variance factors in the wavelength direction (z_j) is generated by using a "double-sigmoidal" wavelength-dependent function with a value near to one near the centre of the spectral region and values of r_{max} at the limits. First a baseline factor (z_0) is selected as 5% of the maximum among all peaks of the pure component spectra. The variance factor corresponding to a wavelength λ is obtained as

$$z_j(\lambda) = \left[1 + (r_{max} - 1)\left(\frac{1}{1 + e^{a(\lambda - \lambda_1)}} + \frac{1}{1 + e^{a(\lambda_2 - \lambda)}}\right)\right] z_0 \tag{19}$$

The value of r_{max} was chosen as 50. The parameters λ_1 and λ_2 are the left and right inflection points, respectively, of the double sigmoid function while a is the slope of the sigmoidal curves given by:

$$a = \frac{4.394}{\Delta\lambda} \tag{20}$$

where $\Delta\lambda$ is the 10%-90% rise range of the sigmoid. The values of λ_1, λ_2, and $\Delta\lambda$ are chosen as 430, 560, and 26, respectively.

The error standard deviation matrix of dimension 20 × 41 are generated by multiplying 20×1 vector of x_i factors and 1 ×41 vector of z_j factors and taking square root of each element. A random number is generated from a $\mathcal{N}(0, 1)$ distribution and then multiplied with each σ_{ij} to obtain the errors. The noisy absorbance measurements are obtained by adding the errors to the true absorbances.

In data set 2, the standard deviation of the error in each mixture is assumed to be wavelength-dependent as given by the function

$$\sigma(\lambda) = \left[1 + (r_{max} - 1)\left(\frac{1}{1 + e^{a(\lambda - \lambda_1)}} + \frac{1}{1 + e^{a(\lambda_2 - \lambda)}}\right)\right] \sigma_0 \tag{21}$$

with a is given by Eq. (20) and r_{max}, λ_1, λ_2, and $\Delta\lambda$ chosen equal to 10, 450, 570, and 24. The value of σ_0 is taken as 5% of the maximum true absorbance among all mixtures.

In data set 3, the standard deviation of error in the absorbance of a mixture at a particular wavelength is taken to be equal to 5 % of corresponding true absorbance. It may be noted that the errors variances in this case do

not follow the factored noise model. It is assumed that the concentration measurements are error free in all three data sets.

Data sets 1-3 assume that the concentration measurements do not contain any error. However, this is not a case with real data sets. Hence, data set 4 analyzes the effect of errors in concentration measurements. Errors in absorbance measurements follows the noise model in data set 2 with r_{max}=3 and σ_0 as 3% of the maximum absorbance in the true data. The standard deviation of error in the concentration measurement is taken to be equal to 5% of the corresponding true concentration. A similar data set with standard deviation of errors equal to 20% of the true concentrations is also generated. Random errors from the gaussian distribution with these standard deviations are then added to the true concentrations to obtain the noisy concentration measurements.

3.2. Experimental data set

Two experimental data sets are also used to test the performance of the proposed method. The first experimental data set (data set 5) contains the absorbances of mixture of three metal ions (Co(II), Cr(III), and Ni(II)) prepared in 4 % HNO_3 solution. This data set was obtained through carefully designed experiments by Wentzell *et al.* [5]. The data set contains absorbance measurements for 26 mixtures between the range of 350 nm and 650 nm at intervals of 2 nm. Five replicate measurements of its absorbance spectra have been made for each mixture. The noise levels near the ends of the wavelength range were increased by using a band-pass filter. The standard deviations of errors at different wavelengths can be estimated directly using the five replicates for each mixture. The spectra and standard deviations estimated from the replicate experiments for this data set are shown in Fig. 1. It can be observed from Fig. 1b that the standard deviation of errors varies with respect to both wavelength and mixtures, but may not necessarily follow the factored noise model.

The second experimental data (data set 6) contains the absorbances of three-component mixtures containing toluene, chlorobenzene, and heptane. The data contains absorbance measurements for 31 mixtures between the range of 400-2500 nm at intervals of 2 nm from an augmented three-level, three-factor factorial design [5]. The concentrations varies between 20 and 70 wt % for toluene and chlorobenzene and between 2 and 10 wt % for heptane. Fig. 2 shows a typical spectrum over the full range and standard deviations obtained from replicate scans. Fig. 2 shows that the certain regions of the

spectrum above 1600 nm have high signal-to-noise ratio and specially above 2000 nm. Hence, the absorbance measurements between the range 400-2000 nm is used in this paper for the development of calibration model.

3.2.1. Computational Aspects

All the simulations on the simulated and experimental data sets are performed in Matlab 7. The computations are performed on a Mac desktop with 1.86 GHz Dual-core intel processor and 1 GB of memory.

4. Results and discussion

4.1. Comparison methodology

The predictive ability of the MVC models is an important aspect when different MVC models are compared for the same data set. Here, the developed MVC models are compared with two methods. First, the conventional cross-validation method *leave-one-score-out* method is used to compare MVC models. A rigorous statistical test, called the randomization test, is then applied to compare the models [8]. Both these methods are briefly described below.

Leave-one-score-out method

The *leave-one-score-out* methodology of cross-validation is used to validate the predictive ability of the calibration model constructed [5]. In this method, the true data subspace is estimated using all the mixtures but the development of the calibration model is carried out by excluding the scores of one of the mixtures. The developed calibration model is used to predict the concentrations of the mixtures whose scores were excluded. This procedure is repeated such that each mixture is excluded once. The root-mean-square error (RMSE) between predicted and actual concentrations is calculated by

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{(c_i^{pred} - c_i^{ref})^2}{m}} \tag{22}$$

where c_i^{pred} and c_i^{ref} are the predicted and reference concentrations of the species in the excluded mixture, respectively, and m is the number of mixtures in the calibration set. The total root-mean-square error (RMSET) is calculated in a similar manner over the concentrations of all the components present in the mixture. The RMSE and RMSET values give an indication of the predictive ability of the calibration model.

Randomization test

A randomization test is a data-driven approach based on random assignment. A detailed description of the method is given in Edgington [10] and van der Voet, H. [9]. Here, the test often called, randomization t-test, is briefly described for a comparison between models Q and R. Let $\mathbf{e}_Q(k)$ and $\mathbf{e}_R(k)$ be the m-dimensional vectors of the predictive errors and the mean squared error of prediction (MSEP) for kth species (k=1, 2, 3) defined as follows:

$$\begin{aligned} \mathbf{e}_Q(k) &= \mathbf{c}_Q^{pred}(k) - \mathbf{c}^{ref}(k),\ \mathrm{MSEP}_Q(k) = \frac{\sum_1^m \mathbf{e}_Q^2(k)}{m}, \\ \mathbf{e}_R(k) &= \mathbf{c}_R^{pred}(k) - \mathbf{c}^{ref}(k),\ \mathrm{MSEP}_R(k) = \frac{\sum_1^m \mathbf{e}_R^2(k)}{m}. \end{aligned} \tag{23}$$

The difference of predictive errors for kth species, $\mathbf{d}(k)$ (of dimension $m \times 1$) and the mean of difference, $\bar{d}(k)$, are computed as follows:

$$\begin{aligned} \mathbf{d}(k) &= \mathbf{e}_Q(k)^2 - \mathbf{e}_R(k)^k \\ \bar{d}(k) &= \sum_1^m \mathbf{d}(k)/m = \mathrm{MSEP}_Q(k) - \mathrm{MSEP}_R(k). \end{aligned} \tag{24}$$

A comparison between models Q and R is made using the a test statistic as follows [10] :

$$T = \sum_{k=1}^{3} \bar{d}(k) \tag{25}$$

Here, the randomization t-test for one-sided alternative hypothesis $\mathrm{MSEP}_Q(k)$ $>\mathrm{MSEP}_R(k)$ (for $k = 1, 2, 3$) is considered and the algorithm for the test is given in van der Voet, H. [9].

4.2. Comparison of performance on simulated data sets

The predictive ability of the calibration model developed by PCR, IWPCR, and MLPCR is evaluated on simulated data sets in this section. The simulated data sets are of dimension 20×41. In order to apply IWPCA, the total number of error variance factors required to estimate for all data sets are equal to $20 + 41 = 61$. The maximum number of error variances that can be estimated is equal to $(17 \times 18)/2 = 153$ for these data sets, and hence, it is theoretically possible to estimate all the factors by apply IWPCA to these data set. However, an attempt to estimate the error variance factors in both directions simultaneously failed due to numerical convergence problems

in the optimization step. Thus, the sub-optimal method described in Section 2.3.1 has been employed to estimate all the error variance factors. First, IPCA is applied to estimate the factors along the sample direction (i.e. x_i). In the next step, IWPCA is applied to estimate the factors along the wavelength direction by keeping the variance factors along the sample direction to be fixed at the estimates obtained using IPCA. Once the variance factors in the sample and wavelength directions are estimated, the true absorbance subspace is estimated using SWPCA method. A lower bound of 0.003 on the estimated error variance factors was also imposed in the sub-optimal method to obtain convergence for all simulated data sets. Following the estimation of the absorbance subspace, the MVC model is obtained using IWPCR method as described earlier.

The results for the first three simulated data sets are presented in Table 1 in terms of RMSE values obtained by different MVC models. Although, the performance of the methods was assessed for different choices for the number of latent factors, only the results for three latent factors are reported. It was observed that the RMSET values for the PCR, IWPCR, and MLPCR methods showed a significant reduction when the dimension of the true data subspace is correctly chosen as three, after which they remained more or less constant. This is consistent with the fact that the number of species in the mixtures is three and therefore the correct dimension of the true data subspace is also three. The results in Table 1 show that IWPCR and MLPCR perform better than PCR. The performance of MLPCR is only marginally better than that of IWPCR. The randomization t-test is applied to test MVC models for data set 1. The results obtained by the randomization t-test is summarized in Table 2. The randomization t-test reveals that the difference between PCR and IWPCR is significant, however the difference between IWPCR and MLPCR is indicative. The difference between PCR and MLPCR is also significant. Hence, the randomization t-test on data set 1 reaffirms the inferences made based on RMSE values. It should be noted that MLPCR requires knowledge of all the error variances, while IWPCR estimates the error variances from the same data set. Figure 3 shows a comparison of the standard deviations estimated using IWPCR method with the true variances used to simulate one of the mixtures in data set 1 (mixture 4), when number of factors is chosen as three. It is observed from Figure 3 that the estimated standard deviations are quite good. The results clearly indicate the ability of the IWPCA method to simultaneously estimate the error variances as well as the true absorbance subspace accurately.

Data set \ Methods	Species	PCR	IWPCR	MLPCR
Data set 1	A	0.049	0.022	0.015
	B	0.052	0.015	0.013
	C	0.050	0.024	0.023
	Total	0.050	0.021	0.018
Data set 2	A	0.114	0.049	0.044
	B	0.163	0.026	0.024
	C	0.156	0.034	0.030
	Total	0.146	0.037	0.034
Data set 3	A	0.010	0.012	0.006
	B	0.014	0.012	0.011
	C	0.011	0.007	0.006
	Total	0.012	0.010	0.008

Table 2: RMSE value comparison of different MVC models for simulated data sets at the number of factors equal to three.

Comparisons with IWPCR			
	Significance values obtained by randomization t-test		
Prediction method	Data set 1	Data set 2	Data set 3
PCR	0.005	0.005	0.30
MLPCR	0.98	0.19	1
	Comparison between PCR and MLPCR		
MLPCR	0.005	0.005	0.005

Table 3: Application of the randomized test on the simulated data for the number of factors equal to 3. All tests are one-sided with 199 trials.

The results of applying the different MVC method for data set 2 are shown in the second row of Table 1. Although data set 2 does not follow the factorial noise model as in IWPCR, the predictive ability of the MVC model developed by IWPCR is better than PCR and as good as MLPCR. The results of randomization t-test applied on data set 2 is presented in Table 2. The randomization t-test shows that the predictive ability IWPCR and MLPCR are better than PCR. There is also some indication that MLPCR performs marginally better than IWPCR. The standard deviations estimated using IWPCR are compared with the true standard deviations for sample number 4 in Figure 4. It is observed from the figure that the estimated standard deviations are comparable to the actual values.

The results of applying the three MVC methods for data set 3 are shown in the last row Table 1. In this case, RMSE values indicates that the performance of all three methods are almost the same. The randomization t-test shows that the performance of IWPCR is same as PCR and MLPCR, although there is evidence to conclude that MLPCR performs better than PCR. It was found that the signal-to-noise ratio (ratio of standard deviation of true spectra to the standard deviation of errors) for this data set ranged from 16 to 23 for all the mixture spectra. Due to this high signal-to-noise ratio PCR performs as good as IWPCR. The comparison of the estimated standard deviations using IWPCR and the true standard deviations for one of the mixtures is shown in Figure 5. Since the noise does not follow the factored model assumption, the estimates of standard deviations are not as good as for data sets 1 and 2. Nevertheless, the estimates do follow the trend of the true noise variances.

In real data sets, the concentrations are measured with uncertainty and hence, data set 4 is simulated with errors in the concentrations also included. The level of errors in the concentration measurements is taken equal to 0, 5 and 20%. The results of applying the three MVC methods for data set 4 at the three-level of uncertainty are summarized in Tables 3 and 4 for the number of factors equal to three. Data set 4 with 0% errors in the concentrations is identical to data set 2. The results in Tables 3 (RMSE) and 4 (randomization t-test) shows that the performance of IWPCR and MLPCR is better than PCR and the difference between MLPCR and IWPCR is only indicative. Similar kind of results are obtained when the errors in the concentrations are increased to 5%. However, the performance of MLPCR is better than PCR and IWPCR when the errors in the concentrations are increased to 20% (see Tables 3 and 4). The RMSE values and randomization t-test

Data set \ Methods	Species	PCR	IWPCR	MLPCR
Data set 4a (without error in concentration)	A	0.093	0.041	0.039
	B	0.158	0.044	0.036
	C	0.140	0.055	0.050
	Total	0.133	0.047	0.042
Data set 4b (with 5% proportional error in concentration)	A	0.112	0.050	0.051
	B	0.138	0.061	0.042
	C	0.130	0.061	0.042
	Total	0.127	0.054	0.045
Data set 4c (with 20 % proportional error in concentration)	A	0.119	0.140	0.132
	B	0.160	0.150	0.108
	C	0.154	0.140	0.105
	Total	0.145	0.144	0.116

Table 4: RMSE value comparison of different MVC models for simulated data sets with errors in concentrations at the number of factors equal to three.

show that the performance of PCR is as good as IWPCR. Since MLPCR uses knowledge of error covariances, the performance of MLPCR is better than PCR and IWPCR.

4.3. Performance comparison on experimental data

In this section, the results obtained by applying PCR, MLPCR, and IWPCR to the experimental data sets 5 and 6 are presented. Data set 5 contains absorbance spectra of 26 distinct mixtures with five replicates for each mixture. Using these replicates, the standard deviations for each measured spectrum has been estimated by Wentzell *et al.* [5]. The main focus of this work is to handle the data set when replicates of spectra are not available. Therefore, non-replicated measurement data consisting of 26 mixtures, have been used for evaluating all three methods. This data set is constructed by randomly picking one mixture spectra from each of the five replicates of the original data set. The resulting random sets are matrices of dimension 26×151. The maximum number of error variances that can be estimated is equal to $(23 \times 24)/2 = 276$. Since this data set is large, the sub-optimal method described in Section 2.3.1 for IWPCR has been used to estimate the error variances. In this sub-optimal method, first IPCA was applied to estimate

Comparisons with IWPCR			
	Significance values obtained by randomization t-test		
Prediction method	Data set 4a	Data Set 4b	Data set 4c
% errors in concentrations	0 %	5 %	20 %
PCR	0.005	0.005	0.48
MLPCR	0.98	0.92	1
Comparison between PCR and MLPCR			
MLPCR	0.005	0.005	0.015

Table 5: Application of the randomized test on the simulated data with the concentration uncertainty for the number of factors equal to 3. All tests are one-sided with 199 trials.

the variance factors in the mixture direction. Then, the data set of 26 × 151 was divided into three sub-matrices of dimensions 26 × 50, 26 × 50, and 26 × 51. In the next step for each sub-matrix, IWPCA was applied to estimate the variance factors along the wavelength direction keeping the estimates of variance factors in the mixture direction to be fixed at the estimates already obtained. This procedure was repeated for all sub-matrices to estimate all the variance factors along the wavelength direction. A lower bound of 0.0001 on the estimated error variance factors was also imposed in the sub-optimal method to obtain convergence for the experimental data set. Then, the calibration model was developed as described in Section 2.4.

The results obtained by applying all three methods are presented in Tables 5 and 6. Two random sets were used to evaluate these three methods. Since there are three species in the data set, the dimension of the true data subspace should be three. However, the real data can be affected by the factors such as the nonlinearity and offsets and thus the results are given for up to six factors. The RMSE values and randomization t-test show that IWPCR performs better than PCR when the dimension of the true subspace is taken to be three (or more). It shows that incorporation of error variance information into the PCA based model definitely improves the predictive ability of the model. Secondly, the RMSE values show that IWPCR gave marginally better results as compared to MLPCR at three latent factors, although no knowledge of error variances is assumed in IWPCR. The randomization t-test indicates that the difference between MLPCR and IWPCR is only indicative when the number of factors are chosen equal to three. However, MLPCR

Number of latent factors	Species	PCR		IWPCR		MLPCR	
		set 1	set 2	set 1	set 2	set 1	set 2
1	Co	11.52	12.06	13.09	13.18	11.2	11.3
	Cr	3.79	3.73	3.68	3.45	3.18	3.14
	Ni	21.28	20.96	19.84	20.96	24.84	25.01
	Total	14.14	14.13	13.88	14.43	15.86	15.95
2	Co	7.55	9.31	3.50	6.55	7.47	7.56
	Cr	3.84	3.45	0.25	2.91	3.25	3.21
	Ni	18.95	15.01	26.68	21.95	17.77	17.92
	Total	11.98	10.39	15.60	13.33	11.28	11.38
3	Co	7.82	8.40	0.37	0.39	0.37	0.38
	Cr	3.30	3.47	0.11	0.10	0.11	0.10
	Ni	14.18	15.29	0.35	0.51	0.41	0.52
	Total	9.54	10.27	0.30	0.38	0.33	0.38
4	Co	7.22	6.27	0.39	0.35	0.39	0.38
	Cr	3.38	2.33	0.12	0.09	0.11	0.09
	Ni	14.14	9.62	0.36	0.47	0.36	0.44
	Total	9.37	6.76	0.31	0.34	0.31	0.34
5	Co	5.85	6.32	0.39	0.35	0.37	0.32
	Cr	1.73	2.44	0.12	0.09	0.11	0.08
	Ni	7.14	10.07	0.39	0.50	0.36	0.43
	Total	5.42	7.01	0.33	0.35	0.30	0.31
6	Co	5.22	6.64	0.42	0.43	0.32	0.31
	Cr	1.27	2.16	0.13	0.11	0.09	0.08
	Ni	4.69	7.22	0.38	0.43	0.38	0.44
	Total	4.12	5.80	0.34	0.42	0.29	0.31

Table 6: Comparison of different MVC model for experimental data set 5 in RMSE

performs better than IWPCR for the choice of four or higher number of factors. It also indicates that these results show that the factorial error variance model is adequate despite the fact that the experimentally estimated error variances do not necessarily obey this model. These results also show that the sub-optimal approach used for the large data sets is able to give reasonable good estimates of the error variance factors, which in turns leads to an accurate MVC model. The estimated values of standard deviations using the sub-optimal approach and using replicate measurements for the sample number 4 are shown in Figure 6. Figure 6 shows that the sub-optimal method gives good estimates of standard deviations of errors.

Data set 6 contains absorbance data corresponding to visible and short-wavelength NIR range with 31 samples at 1050 wavelengths (400-2500 nm). However, as mentioned in Section 3.2, the data set above 2000 nm is discarded and hence, the truncated data is of dimension (31×800). For estimating the variance factors along the wavelength direction in the IWPCA method, the sub-optimal method is applied by dividing the data set into sub-matrices of dimensions 31×25. The variance factors along wavelength direction are first

Comparisons with IWPCR				
	Significance values obtained by randomization t-test			
# Factors / Prediction method	3	4	5	6
PCR	0.005	0.005	0.005	0.005
MLPCR	0.035	0.99	0.87	1
	Comparison between PCR and MLPCR			
MLPCR	0.005	0.005	0.005	0.005

Table 7: Application of the randomized test on the experimental data set 5 (set 1) for the number of factors equal 3 to 6. All tests are one-sided with 199 trials.

estimated followed by estimating the variance factors along mixture direction. The results obtained by applying all three methods for RMSE values and the randomization t-test are given in Tables 7 and 8. The RMSE values in Table 7 indicate that PCR performs poorly for all the number of factors. In case of IWPCR, it performs poorly for the number of factors equal to three. However, when the number of factors are chosen equal to four or five, there is a marginal difference between the RMSE values obtained by IWPCR and MLPCR. The poor performance of IWPCR for the number of factors equal to three could either be due the presence of a baseline spectra or due to the error variances not following the factored noise model. In spite of marginal difference in RMSE values of MLPCR and IWPCR, the randomization t-test in Table 8 shows that MLPCR performs better than IWPCR. In case of PCR, it performs as good as IWPCR for the number of factors equal to three but performs poorly when number of factors equal to four or five.

In all the simulated and experimental data sets, the true rank of the measured absorbance matrix (dimension of the absorbance subspace) may not be known *a priori* and may have to be estimated. Typically, PCA methods estimate the true rank by examining the singular values. An approach is to look for a sharp change in the singular values by using a SCREE plot [7]. However, for many practical data sets the SCREE plot may not reveal sharp changes. On the other hand, if the noise follows the model used to develop IPCA or IWPCA, then it was shown that the lowest rejected singular values should all be close to unity, while the other singular values corresponding to the chosen rank of the subspace (number of factors chosen) should be greater

Number of latent factors	Species	PCR	IWPCR	MLPCR
3	Toluene	5.00	6.34	0.12
	Chlorobenzene	6.27	10.4	0.13
	Heptane	2.94	2.79	0.09
	Total	4.93	7.36	0.11
4	Toluene	2.66	0.25	0.13
	Chlorobenzene	6.29	0.12	0.10
	Heptane	2.46	0.08	0.07
	Total	4.19	0.17	0.10
5	Toluene	2.77	0.19	0.13
	Chlorobenzene	6.51	0.11	0.11
	Heptane	2.50	0.07	0.07
	Total	4.33	0.14	0.10

Table 8: RMSE value comparison of different MVC model for experimental data set 6

Comparisons with IWPCR			
	Significance values obtained by randomization t-test		
# Factors / Prediction method	3	4	5
PCR	1	0.005	0.005
MLPCR	1	1	1
	Comparison between PCR and MLPCR		
MLPCR	0.005	0.005	0.005

Table 9: Application of the randomized test on the experimental data set 6 for the number of factors equal 3 to 5. All tests are one-sided with 199 trials.

than unity. Figure 7 shows the plot of logarithm of singular values obtained at convergence of the IWPCA method for different number of factors chosen for the experimental data set. Figure 7 shows that the theoretical result may not be precisely satisfied (since the errors may not satisfy the assumption of factored noise model). However, if we closely examine the singular values for different factors in Figure 7, it is observed that if 3 or 4 factors are chosen, the corresponding 3 or 4 singular values are greater than unity while the remaining singular values are reasonably close to unity. But in the case of 5 and 6 factors, more than 5 and 6 singular values are greater than unity. Furthermore, the singular values also show a sharp decrease after the first four singular values. From these observations, the true rank for the experimental data set may be estimated as three or four. Another practical measure for determining the dimension of the true data subspace is to look for a sharp decrease in the RMSET values of the MVC models developed as the number of factors chosen is increased. It is observed from the results presented in all tables, that in the case of IWPCR and MLPCR for all data sets the RMSE and RMSET values decrease significantly when the number of factors is chosen as three after which the change is insignificant. Thus, it is easy to conclude that the true rank of the absorbance matrix is three for these two methods. However, using PCR such a sharp change is not observed for data set 2 and the experimental data set, and it is not easy to accurately estimate the true rank of the absorbance matrix. These results further confirm that the use of either estimated or known error variances to scale the data is necessary and useful to estimate the correct number of independent species and to develop accurate MVC models.

5. Conclusion

The focus of this work was to develop multivariate calibration models for non-replicated measurements. A new method called IWPCA was developed for simultaneously estimating a true subspace (model) and error covariance matrix from the data matrix for the factorial noise model. The MVC method called IWPCR developed using the proposed approach was found to provide more accurate estimates than PCR models on both simulated and experimental data sets. IWPCR also gives as good performance as MLPCR without the requirement of replicate measurements and *a priori* information on error variances when the data set follows factorial noise model. A sub-optimal approach for applying IWPCR to large data sets is also proposed which al-

lows large number of error variance factors to be estimated. These attractive features of the IWPCR method makes it a practically useful method for developing accurate MVC models and in other applications where PCA is used.

References

[1] Narasimhan, S. and Shah, S. L., *Control Engineering Practice* , **16**, (2008), 146-155.

[2] Bhatt, N. P., Mitna, A. and Narasimhan, S., *Chemom. Intell. Lab. Syst.*, **85**, (2007), 70-81.

[3] Paatero, P. and Tapper, U., *Chemom. Intell. Lab. Syst.*, **18**, (1993), 183-194.

[4] Cochran, R. N. and Horne, F. H., *Anal. Chem.*, **46**(6), (1977), 846-853.

[5] Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K. and Kowalski, B. R., *J. Chemom.*, **11**(4), (1997*a*), 339-366.

[6] Wentzell, P. D., Andrews, D. T. and Kowalski, B. R., *Anal. Chem.*, **69**(13), (1997*b*), 2299-2311.

[7] Jackson, J.E. , *A User's Guide to Principal Components*, John Wiley & Sons, New York, (2003).

[8] Faber, N. M. and Rajkó, R., *Anal. Chim. Acta.*, **595**, (2007), 98-106.

[9] van der Voet, H., *Chemom. Intell. Lab. Syst.*, **25**, (1994), 313-323.

[10] Edgington, S.E. , *Randomization tests*, Marcel Dekker, Inc., New York, 2nd edn., (1987).

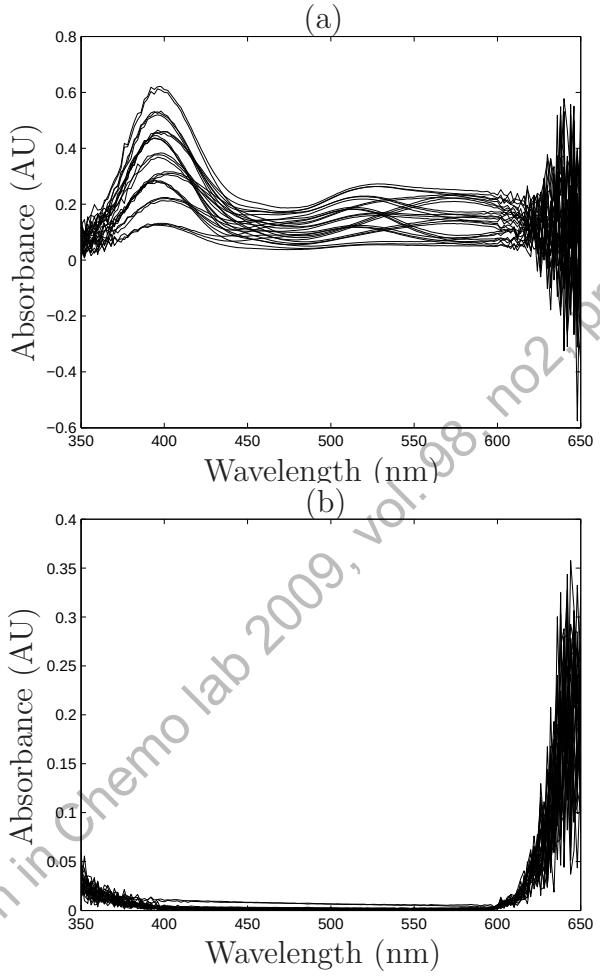

Figure 1: Experimental data set 5: (a) spectra for metal ion mixtures (b) standard deviations of measurement errors.