MISSING DATA TREATMENT USING ITERATIVE PCA AND DATA RECONCILIATION

S.A. Imtiaz [*] **S. L. Shah** [*,1] **S. Narasimhan** [**]

[*] *Department of Chemical and Materials Engineering*
University of Alberta, Edmonton AB, Canada, T6G 2G6
[**] *Department of Chemical Engineering*
IIT,Madras

Abstract: Two methods, one based on Iterative Principal Components Analysis(IPCA) and the other based on Data Reconciliation have been developed for estimating a model from a data matrix containing missing data. These algorithms are iterative in nature and analogous to the method based on PCA for treating missing data. The methods incorporate information about the measurement errors to develop the models and are optimal in a maximum likelihood sense. The close connection of the methods with the Expectation Maximization (EM) algorithm is also established. Simulated data from a Flow Network system with a variety of error structures and missing data is used to evaluate the performance of the proposed methods. In all cases, models estimated by the proposed methods were superior to those obtained by the classical PCA-based missing data treatment algorithms for nonuniform error.

Keywords: Missing Data, PCA, IPCA, Data Reconciliation, Expectation Maximization (EM)

1. INTRODUCTION

Process industries are now using multivariate statistical methods frequently along side with univariate methods. As a result treatment of missing data has become more important from statistical process control (SPC) perspective. Unless we have an appropriate method for dealing with missing measurements, one may end deleting an entire row of data even if a single measurement in that row is missing. When a process is operating at quasi-steady state or in dynamic mode the samples have time stamps and deletion of rows due to missing or bad data leads to irregular sampling intervals. Most of the popular model building techniques such as Principal Components Analysis (PCA), Partial Least Squares (PLS) were originally developed for dealing with uniformly sampled data without any missing value. Modifications are necessary to accommodate missing values in the data matrix. In this paper our focus is on Principal Components Analysis only. PCA is a widely used method to build models from a data matrix of large number of highly correlated variables. Models developed using PCA are subsequently used to monitor process operating performance and detect impending faults. In dealing with data matrices containing missing values, PCA faces difficulties at two stages. Firstly, in building model from data sets with missing values and, secondly in monitoring phase when model is available from previous analysis but some of the monitored variables have missing values. Both problems have received considerable attention from researchers. Historically, NIPALS algorithm was used originally for build-

[1] author to whom all correspondence should be addressed. E-mail: sirish.shah@ualberta.ca

ing principal component models. Christofferson (1970) extended NIPALS algorithm for finding first and second principal component in presence of missing values in the data matrix (Grung and Manne, 1998). Now it has become more customary to use Singular Value Decomposition (SVD) algorithms for finding the PCs. In the presence of missing values in the data matrix, an iterative imputation approach is used to fill the missing values. Initially, the missing values are filled with the unconditional mean of the respective variable and SVD is done on the complete data set. PCA solution is iteratively improved by updating the missing values with the estimated values of the current model until the imputed values stabilize (Grung and Manne, 1998; Troyanskaya *et al.*, 2001; Walczak and Massart, 2001). Assuming that the model is available from a previously obtained data without any missing value Nelson et al.(1996) and Arteaga et al.(2001) developed and analyzed several methods for calculating scores from a data set with missing values. These methods such as, Trimmed Score (TRI), Single Component Projection (SCP), Conditional Mean Replacement (CMR), Projection to Model Plane (PMR) are useful from a monitoring point of view. However all these methods use PCA in its natural form (Nelson and MacGregor, 1996; Artega and Ferrer, 2002). Though PCA is a robust and widely used tool for model building but it has some limitations too. As such, algorithms based on PCA for building models from data matrix with missing values also suffer from these limitations. In this paper we will point out two such restrictions and subsequently propose two algorithms which outperform currently used PCA-based missing data treatment techniques in these situations.

Firstly, PCA is often not the optimal method for model estimation in Maximum Likelihood sense. For example, ordinary least squares is a maximum likelihood method when measurement errors in the regressor variables are negligible and error in the dependent variable are independent and identically distributed. Likewise, PCA can be considered maximum likelihood if all the measurement errors are uncorrelated and of equal variance (i.e. independent and identically distributed, iid). This is a very restrictive assumption as process measurements come from a wide variety of sensors (i.e. thermocouple, gas chromatograph etc.) and these different types of sensors have different measurement accuracy. Sometimes instead of using these raw variables, derived variables are used in the data matrix, for example enthalpy which is a derived variable obtained by multiplying flowrate with temperature and heat capacity. Thus, there is a possibility that errors may be cross correlated as well. Wentzell et al.(1997) developed Maximum Likelihood Principal Component Analysis (MLPCA) which relaxed these assumptions and gives ML estimate of the model using alternating regression method. MLPCA requires a priori knowledge of the true values of the error covariance matrix (Wentzell *et al.*, 1997). In chemometrics applications often replicates of the measurements are taken and an estimate of the error covariance is available. Whereas, in process industries replicates of measurements are rarely taken. Recently, Narasimhan and Shah (2004) developed an iterative technique, Iterative Principal Components Analysis (IPCA) which estimates both the model and the error covariance matrix simultaneously (Narasimhan and Shah, 2004). In this paper we will show the use of IPCA in missing data context.

The other difficulty associated with PCA is to decide on the number of principal components to be included in the model. PCA relies on ad hoc. methods such as, scree test, cross validation etc. These methods are not backed by theory and they often lead to inaccurate model order selection. The problem becomes even more serious in the presence of missing values in the data matrix. IPCA provides an accurate way of determining the number of independent variables by looking at the singular values.

Two algorithms, one based on IPCA and the other based on IPCA and Data Reconciliation have been proposed in this paper for treating missing data. Both methods are iterative imputation type and henceforth labelled as IPCA Imputation Algorithm (IPCAIA) and IPCA Data Reconciliation (IPCADR). The algorithms are optimal in the Maximum Likelihood Sense for steady state and quasi-steady state processes. Several assumptions are inherent in the development of the methods. First, a true underlying model of lower dimension than the total number of variables does exist, Second, deviations of the measurements are only due to random measurement errors, Third, measurement errors are normally distributed and Finally, the methods are only valid for treating missing data which are Missing Completely At Random (MCAR) or Missing At Random (MAR). Treatment of missing data with Non Ignorable (NI) mechanism is beyond the scope of the present paper (Little and Rubin, 2002).

The paper is organized as follows, in section 2 IPCA algorithm and theory behind the algorithm have been described. Missing data treatment algorithm using IPCA is introduced in section 3. Section 4 gives a brief overview of data reconciliation and shows its application in combination with IPCA algorithm for handling missing data. In section 5 improved performance of these newly developed algorithms has been demonstrated by a flow network simulation example.

2. ITERATIVE PRINCIPAL COMPONENT ANALYSIS (IPCA)

The details of IPCA can be found in Narasimhan and Shah(2004). Here only the main steps of IPCA have been discussed for completeness. IPCA has two main steps, first, Optimally Scaled Principal Components Analysis (OSPCA) and second, Estimation of the error covariance matrix. IPCA iteratively alternates between these two steps until convergence.

An optimal scaling strategy has been used in IPCA. It was shown that PCA is scaling invariant with this scaling scheme. Let us consider a case when all the measurements contain random errors and are described by

$$Y = X + \varepsilon$$

where $\varepsilon \sim N(0, \Sigma_\varepsilon)$ and the underlying model is such that,

$$AX = 0$$

Scaling factor L is defined by,

$$LL^T = \Sigma_\varepsilon$$

After scaling the measurements with the scaling matrix the transformed measurements are given by,

$$Y_s = L^{-1}X + L^{-1}\varepsilon$$

Covariance of Y_s is given by,

$$\Sigma_{y_s} = S_{x_s} + I$$

This is an important result as it provides a convenient way to select the order of the model. From the *eigenvalue* shift theorem, eigenvectors of noise corrupted covariance matrix, Σ_{y_s} is equal to the eigenvectors of the noise free covariance matrix, S_{x_s}. So there is no distortion in the eigenvectors due to the presence of noise in the signals. On the other hand, eigenvalues of the noise corrupted covariance matrix are shifted by unity from the noise free eigenvalues. For example, if the rank of the data matrix X_s is m, then the last n-m eigenvalues of S_{x_s} will be exactly zero as the data matrix is noise free and the last n-$m$eigenvalues of Σ_{y_s} will be unity. The eigenvectors corresponding to these unity eigenvalues define the basis vectors of the residual space, which is the constraint model in scaled domain, A_s. Therefore, the model order selection is not arbitrary. Rather it provides a definitive way of selecting the model order. The constraint model in the original domain is simply given by,

$$A = A_s L$$

The only unknown parameter is Σ_ε. To estimate Σ_ε an iterative approach is taken. An initial estimate of the constraint model, A^0 is obtained by ordinary PCA on the unscaled data matrix Y. Using this initial estimate residuals at each instant are calculated as, $r(t) = A^0 y(t)$

If the estimated model is exact, then the residuals will be independent normally distributed variables with zero mean. Thus the joint density function of *r(1) r(2) r(3)....r(N)* is easily obtained and the maximum likelihood estimate of Σ_ε is estimated by maximizing the log likelihood function of *r(1) r(2) r(3)....r(N)*, which is equivalent to minimizing the following function:

$$\min_{\Sigma_\varepsilon} N log\left|\hat{A}^0\Sigma_\varepsilon\left(\hat{A}^0\right)^T\right| + \sum_{i=1}^{N}\left(r_i^T(t)\left(\hat{A}^0\Sigma_\varepsilon\left(\hat{A}^0\right)^T\right)^{-1} r_i(t)\right) \quad (1)$$

The maximum number of elements in Σ_{y_s} that can be estimated is restricted by the number of constraints or the rank of A. If rank of A is m then the maximum number of elements that can be estimated is less than or equal to *m(m+1)/2*. The method only allows error covariance in the spatial direction.

3. IPCA IMPUTATION ALGORITHM (IPCAIA)

IPCA algorithm has been discussed in the previous section. In this section we will describe the main computational steps of IPCAIA and show its link to the much celebrated Expectation Maximization (EM) Algorithm.

Let $Y^{n\times N}$ be the data matrix containing the noise corrupted measurements which include both dependent and independent variables, where n is the number of variables and N is the total number of samples. If Y_{obs} and Y_{mis} denote the observed and the missing values respectively then the data matrix, $Y = \{Y_{obs}, Y_{mis}\}$.

$$Y = X + \varepsilon$$

where $X$is the noise free variables and ε is measurement noise. There is no assumption on the distribution of the variables Y or X. Only assumption is that, ε follows a multivariate normal distribution with mean zero and covariance Σ_ε, and the underlying model is such that,

$$AX = 0$$

The main steps of IPCAIA are given below:

(1) The missing values of the data matrix are filled with the unconditional mean of the variables. For example, the missing values of the data matrix are filled by the row averages of Y_{obs}.
(2) The filled data matrix Y is supplied to the IPCA algorithm. IPCA automatically determines the number of significant principal components and gives an estimate of the constraint matrix A, scaling matrix L and predicts the noise free variables in scaled domain, $\hat{X}_s$.

(3) The estimated noise free variables $\hat{X}_s$ are converted to $\hat{X}$ in the original domain, $\hat{X} = L\hat{X}_s$. Missing values in the data matrix are filled with these predicted $\hat{X}$ values.
(4) Return to step (2) and repeat these steps until convergence.

The link between IPCA and Expectation Maximization (EM) algorithm is clear from the iterative nature of the two algorithms. Each iteration of EM consists of an Expectation step and a Maximization step. The E-step calculates the conditional expectation of the sufficient statistics given the observed data and the current estimated parameters. Because of filling the missing values with the conditional expectation the variances are underestimated and the main feature of the E-step is adding corrections to the variances. Since error variances are estimated in IPCAIA this can be done implicitly by adding a scaled error term to each of the estimates of the missing value. In step 3 of the above algorithm, instead of filling the missing values with $\hat{X}$ missing values may be filled with $\hat{Y}$ where,

$$\hat{Y} = \hat{X} + L\nu$$
$$\nu \sim N(0, I)$$

The estimated conditional expectation of the sufficient statistics from the E-step is substituted in the log likelihood function and parameters of the model are estimated by maximizing the function. This is the M-step of EM algorithm. Important fact is that parameter estimation has to be maximum likelihood estimate for the observed data set(Little and Rubin, 2002).

In IPCAIA parameter estimation using IPCA is equivalent to the maximization step of the EM algorithm. The parameters of the model are obtained by minimizing the objective function,

$$\underbrace{\min}_{A_s, X_s} \sum_{i=1}^{n} (Y_{s,i} - X_{s,i}) (Y_{s,i} - X_{s,i})'$$

Because of the optimal scaling strategy, in the scaled domain the error covariance matrix is iid. Thus, IPCA gives an ML estimate of the model parameters under this optimal scaling scheme. Since the missing values are filled by values from the distribution of the data, the method remains optimal in the presence of missing data.

4. IPCA DATA RECONCILIATION (IPCADR)

The IPCA Data Reconciliation algorithm for treating missing data is a combination of two methods: IPCA and Data Reconciliation. Before describing the actual algorithm a brief introduction on Data Reconciliation is necessary. In presence of measurement error, the balance equations are not satisfied exactly. The objective of data reconciliation is to estimate the underlying noise free variables so that they satisfy the balance equations. In accordance with the earlier notation,

$$y(t) = x(t) + \varepsilon(t)$$

where $y(t)$ is a nx1 vector of observations, $x(t)$ is a nx1 vector of the underlying variables and $\varepsilon(t)$ normally distributed measurement error. If the time invariant constraint model is A then data reconciliation solves the following weighted least squares problem:

$$\begin{aligned} &\underset{x}{min}\, (y - x)^T \Sigma_{\varepsilon}^{-1} (y - x) \\ &s.t. Ax = 0 \end{aligned} \tag{2}$$

If all the measurements are not available then the x values corresponding to these unmeasured variables are estimated using the constraints and the available measurements. As the number of unmeasured values increases the degree of freedom also increases, so the estimation becomes poor. In extreme cases when there is not enough redundancy between the measurements and the model, data reconciliation breaks down(Sanchez and Romagnoli, 1996). This is important with regard to missing data treatment because the degree of redundancy changes with missing measurement. So, while using data reconciliation for filling missing data we expect that the method will perform better for small number of missing values in the data matrix.

The main steps of the IPCADR algorithm are as follows:

(1) The missing values of the data matrix are filled with the unconditional mean of the variables. For example, the missing values of the data matrix are filled by the row averages of Y_{obs}.
(2) The filled data matrix Y is supplied to IPCA algorithm. IPCA automatically determines the number of significant principal components and gives an estimate of the constraint matrix A and the error covariance matrix Σ_{ε}.
(3) The estimated constraint matrix and the error covariance matrix are used in equation (2) to estimate the underlying variables $x(t)$ at each instant, where t=1.....N. If enough redundancy is not present in the data then the missing values are filled by the unconditional mean.
(4) Return to step(2) and repeat these steps until convergence.

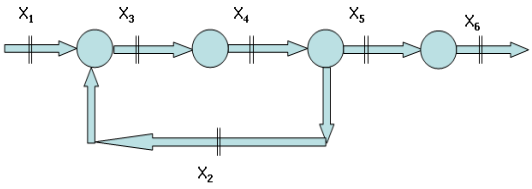

Fig. 1. *Schematic Diagram of the Flow Network*

Table 1. Transfer Functions of the Deterministic Signals

X_1	X_2
$\frac{1}{0.9s+1}$	$\frac{1}{0.8s+1}$

5. SIMULATION RESULTS AND DISCUSSIONS

A flow network process, shown in figure 1, has been chosen to demonstrate the strength of the newly proposed methods. It is assumed that the fluid flowing through the network is incompressible and there is no time delay in the process. The constraint model, A of the process can be obtained easily from the mass balance equation at the junctions. Four mass balance equations can be written for this flownetwork system, so the model order of the constraint matrix is four. In the above example X_1 and X_2 were chosen as independent variables. These are two deterministic signals with slow dynamics. Transfer functions used to simulate these two flows are given in table 1. The rest of the flow rates, X_3 to X_6 were estimated from the mass balance equations. In accordance with the previous notations these are the noise free variables and satisfy the model,

$$AX = 0$$

The observed flow rates are corrupted by measurement noise only,

$$Y = X + \varepsilon$$

Measurement noise were assumed uncorrelated but with unequal variances $\left(i.e. \varepsilon \sim N\left(0, \sigma_i^2 I\right)\right)$. Each measurement vector has 200 data points.

Two indicators were used to assess the performance of the proposed methods. First, Total Sum of Squared Error (TSE), which is defined as,

$$TSE = \sum_{j=1}^{N} \sum_{i=1}^{n} \left(X_{ij}^{mis} - \hat{X}_{ij}^{mis}\right)^2$$

Secondly, Subspace Angle(θ), this is the angle between the two subspaces specified by their columns. If the true constraint model is A^{6x4} and the estimated constraint model is $\hat{A}^{6x4}$, then the angle between each column is defined by,

$$\theta_i = \left\| A_{.i} - A_{.i}\hat{A}^T \left(\hat{A}\hat{A}^T\right)^{-1} \hat{A} \right\|$$

$$\theta = \sum \theta_i$$

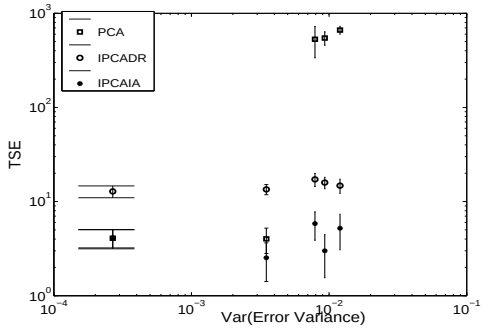

Fig. 2. *Prediction Capability Comparison. Total Sum Square Error vs. variance of error variances*

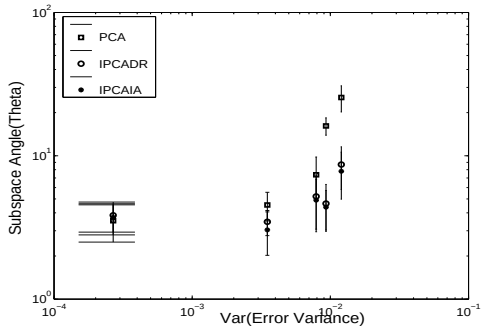

Fig. 3. *Model Quality Comparison. Subspace Angle between the estimated model and actual model vs. variance of error variances*

For a two dimensional space this is the angle at the intersection of two planes. If the angle is small, the two matrices will be nearly linearly dependent.

It was claimed in the previous sections that the proposed methods perform better than PCA based imputation algorithms in two situations, first, model estimation when the errors are not iid $\left(i.e. \Sigma_\varepsilon \neq \sigma^2 I\right)$ and second, model order selection in the presence of missing data. To demonstrate the first point, the Total Sum Square Error (TSE) and the Subspace Angle (θ) estimated by all three methods are plotted in figure 2 and 3. The deviations of the noise variance from iid was quantified by var(error variance). In both cases 7.3% of the total values were missing. TSE gives a measure of the prediction capability of the algorithms. When the noise covariance is close to iid performance of all three methods are practically indistinguishable. But as it deviates further from the iid assumption the performance of the PCA based algorithm deteriorates sharply. The subspace angle is also smaller for the models estimated by IPCAIA and IPCADR algorithms compared to the PCA based algorithm.

The results on the robustness of the algorithms are shown in figure 4. In this case the error covariance matrix deviates from iid only moderately; var(error variance) is 0.0035. As the missing data increases above 10% of the total data, the difference between the estimated models becomes apparent. Estimated model using IPCAIA algorithm

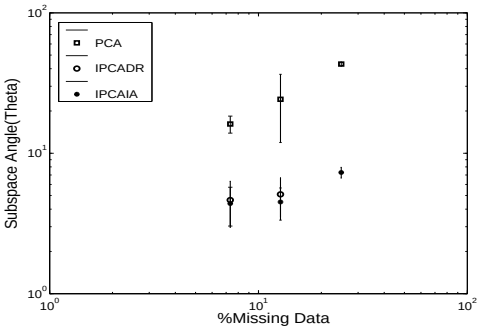

Fig. 4. *Robustness Comparison Plot. Subspace Angle between the estimated model and actual model vs. % missing data*

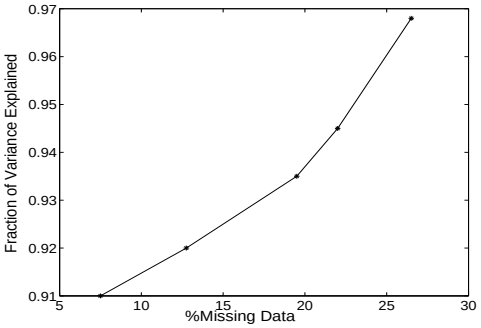

Fig. 5. *Fraction of Variance Explained vs. %Missing Value*

is very good for missing values as high as 25%, while the PCA-based algorithm performs poorly for this large amount of missing data. On the other hand, IPCADR algorithm performs well up to moderate range of missing data. But at a high range the method breaks down because of lack of redundancy.

The improved model order selection criteria in the presence of missing data are evident from figures 5 and 6. Figure 5 shows that as more values are missing in the data matrix, total variance explained by first two principal components increases (figure 5). So, methods used in PCA for selecting model order, such as Scree plots suffers even more in the presence of missing data. On the other hand, the last four eigenvalues from IPCAIA and IPCADR do not deviate significantly from unity (figure 6). Therefore, the model order selection is precise for these two methods even in the presence of missing data.

6. CONCLUSIONS

In this work two limitations of PCA in connection with missing data treatment have been pointed out: 1) model order selection in the presence of missing data, 2) model building from data with unequal noise variances. The proposed algorithms IPCAIA and IPCADR provide a precise way of model order selection by looking at the eigenvalues. Error covariance is incorporated in the

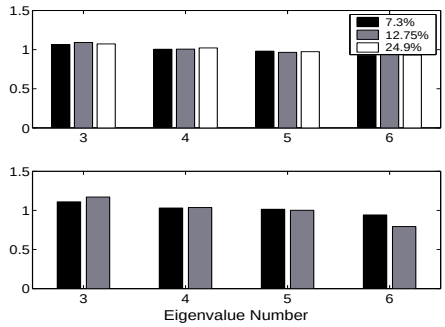

Fig. 6. *Last four eigenvalues estimated by IPCAIA(top) and IPCADR (bottom)*

model estimation procedure. Thus, a maximum likelihood estimate of the model parameters can be obtained. From the simulation study it was evident that IPCAIA was able to estimate good models up to a fairly significant amount of missing data (25%), while IPCADR showed good performance up to moderate range (13%). Both methods outperformed PCA-based missing data treatment algorithm for the unequal noise situation.

REFERENCES

Artega, R. and A. Ferrer (2002). Dealing with missing data in MSPCS: several methods, different interpretations, some examples. *Journal of Chemometrics* **16**, 408–418.

Grung, B. and R. Manne (1998). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **42**, 125–139.

Little, R.J.A. and D.B. Rubin (2002). *Statistical Analysis with Missing Data*. Vol. 2. John Willy and Sons.

Narasimhan, S. and S.L. Shah (2004). Model identification and error covariance matrix estimation from noisy data using PCA. ADCHEM.

Nelson, P.R.C., Taylor P.A. and J.F. MacGregor (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems* **35**, 45–65.

Sanchez, M. and J. Romagnoli (1996). Use of orthogonal transformations in data classification reconciliation. *Computers and Chemical Engineering* **20**(5), 483–493.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinfomatics* **17**(6), 520–525.

Walczak, B. and D.L. Massart (2001). Dealing with missing data. *Chemometrics and Intelligent Laboratory Systems* **58**, 15–27.

Wentzell, P.D., E.T. Andrews, A.D. Hamilton, K. Faber and B.R. Kowalski (1997). Maximum likelihood principal component analysis. *Journal of Chemometrics* **11**, 339–366.