Available online at www.sciencedirect.com
ScienceDirect
Control Engineering Practice 16 (2008) 146–155
CONTROL ENGINEERING PRACTICE
www.elsevier.com/locate/conengprac

Model identification and error covariance matrix estimation from noisy data using PCA

Shankar Narasimhan[1], Sirish L. Shah*

Department of Chemical and Materials Engineering, University of Alberta Edmonton, Alta., Canada T6G 2G6

Received 9 January 2007; accepted 19 April 2007
Available online 27 June 2007

Abstract

Principal components analysis (PCA) is increasingly being used for reducing the dimensionality of multivariate data, process monitoring, model identification, and fault diagnosis. However, in the mode that PCA is currently used, it can be statistically justified only if measurement errors in different variables are assumed to be i.i.d. In this paper, an iterative algorithm for model identification using PCA is developed for the case when measurement errors in different variables are unequal and are correlated. The proposed approach not only gives accurate estimates of both the model and error covariance matrix, but also provides answers to the two important issues of data scaling and model order determination.

Keywords: PCA; Model identification; Measurement errors; Data scaling

1. Introduction

Principal components analysis (PCA) is a multivariate statistical tool developed primarily to obtain a parsimonious representation of multivariate data (Rao, 1964). This is achieved by choosing a few linear combinations known as principal components (PCs), which together capture most of the variability in the data. The number of linear combinations chosen is typically less than the number of measured variables. In chemical engineering, PCA has been used in a similar manner for data compression and for developing soft sensors (Liu, 2007). In recent years, PCA is also gaining significant importance as a tool for model identification or to discover the underlying spatial and/or temporal relationships between variables. For example, PCA is a critical part of many subspace-based dynamic model identification methods (Viberg, 1995). The model identified using PCA has also been subsequently used in fault diagnosis (Yoon & MacGregor, 2000).

If measurements are corrupted by random errors, then PCA is an optimal procedure for estimating the model parameters only if the errors in different variables are independently and identically distributed (Wentzell, Andrews, Hamilton, Faber, & Kowalski, 1997). An improved approach, called the maximum likelihood PCA (MLPCA), has been developed by Wentzell et al. (1997) for general error covariance matrix structures. However, their method assumes that the measurement error covariance matrix is known. It would be advantageous if the measurement error covariance matrix can be estimated along with the model from the same data set. This becomes especially important in chemical processes, since the model as well as the error covariance matrix is likely to change over time. Recently, a method has been proposed for estimating noise variances from non-replicated measurements for a linear regression model (Brauwere, Pintelon, Ridder, Schoukens, & Baeyens, 2007). This method assumes, however, that the regression model is known.

In this paper, an iterative method is developed which combines PCA with a maximum likelihood estimation (MLE) procedure for obtaining an estimate of the error covariance matrix simultaneously along with the process model. The proposed approach also provides answers to

*Corresponding author. Tel.: +1 780 492 5162; fax: +1 780 492 2881.
E-mail address: sirish.shah@ualberta.ca (S.L. Shah).
[1]Department of Chemical Engineering, IIT Madras, Chennai 600036, India.

0967-0661/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.conengprac.2007.04.006

important questions on how to scale measured data before applying PCA, and how to obtain the model order without a priori knowledge.

2. Model identification using PCA with noise free data

The case of model identification using PCA when the measurements are not corrupted by noise is first discussed. Although this is well known, an alternative viewpoint is presented which motivates the development of the proposed approach. The following process identification problem is considered, which, despite its simplicity, contains the essential features for describing more complex processes.

Let $x(t) : n \times 1$ be a vector of n variables at time instant t, which are related by the following set of m independent linear constraints:

$$Ax(t) = 0, \tag{1}$$

where $A : m \times n$ is a constant time invariant constraint matrix. The above equations represent the spatial relations between variables, which are assumed to hold at all time instants. At each time instant, measurements $y(t)$ of all the variables corrupted by random errors are available, which can be written as

$$y(t) = x(t) + \varepsilon(t). \tag{2}$$

The random errors, $\varepsilon(t)$, are assumed to be temporally independent and to follow a multivariate normal distribution with mean zero and covariance matrix Σ_ε.[2] The random errors are also assumed to be independent of $x(t)$. Given a sample of N measurements, $y(1) \ldots y(N)$, the objective is to estimate the constraint matrix. It should be noted that in this paper, the term 'model' is used to refer to the constraint matrix A. In multivariate regression, the term 'model' is used to describe the regression matrix which relates the variables $x(t)$ to a specified set of $n-m$ independent variables from among the n variables. It will be shown, subsequently, that the problem of identifying the regression matrix is the complementary problem to identifying the constraint matrix.

The assumptions that are made regarding the true variables $x(t)$ give rise to two well-known problems in the statistical literature. If it is assumed that $x(t)$ follows a multivariate normal distribution with mean μ and covariance matrix Σ_x, then the problem is referred to as a structural regression problem. On the other hand, if a less restrictive assumption is made that the true values of variables form a fairly arbitrary deterministic sequence, then a functional regression problem is obtained (Fuller, 1987). For chemical processes, if $x(t)$ are considered to be representative of different 'steady-state' operating points, then the assumption of a structural model cannot be justified and therefore the functional model problem is considered in this paper. It should be noted that even for the simple univariate regression problem, it is well known that the structural problem is relatively easy to solve as compared to the functional problem and this difficulty extends to the multivariate regression problem.

Even though $x(t)$ is considered to be a fairly arbitrary deterministic sequence, the following two conditions have to be imposed, which essentially imposes finite limits on its evolution characteristics:

$$\lim_{N\to\infty} \{\sqrt{N}(\bar{x} - \mu_x)\} = 0, \tag{3}$$

$$\lim_{N\to\infty} \sqrt{N}\left[\sum_{i=1}^{N} (x(i) - \mu_x)(x(i) - \mu_x)^{\mathrm{T}} - \Sigma_x\right] = 0, \tag{4}$$

where $\bar{x}$ represents the average of the sequence $x(t)$, and μ_x and Σ_x are bounded.

It can be easily observed that due to the constraints, the vectors $x(t)$ span a $n-m$ dimensional subspace of R^n (denoted as V_x). Furthermore, the rows of A span an m dimensional subspace of R^n (denoted as $V_x^\perp$), which is orthogonal to V_x. Thus, given a sample of measurements in R^n, the objective of model identification can be viewed as the problem of decomposing R^n into two orthogonal subspaces, one of which defines V_x and the other $V_x^\perp$. It can be further noted that in order to define V_x and $V_x^\perp$, a basis for each of these spaces has to be identified. Thus, for identifying the model, it is sufficient to estimate any m linearly independent vectors in the row space of A.

In the absence of measurement errors, if a sample of $n-m$ linearly independent realizations of $x(t)$ are given, then these can be used as a basis for V_x. It is then possible to determine m linearly independent vectors orthogonal to V_x, which define a basis for $V_x^\perp$ exactly. It should be noted that this is sufficient to solve the stated problem.

On the other hand, if PCA is used to solve the above problem, then the orthonormal eigenvectors of the data covariance matrix $S_y = (1/N)YY^{\mathrm{T}}$ are determined, where

$$Y = [y(1), y(2), \ldots, y(N)]. \tag{5}$$

In the absence of measurement errors, S_y is identical to $S_x = (1/N)XX^{\mathrm{T}}$. Therefore, the column space of S_y is identical to V_x, and the matrix S_y has rank $n-m$. Thus, it will have $n-m$ nonzero eigenvalues, while the rest are zeros. The eigenvectors corresponding to the nonzero eigenvalues is an orthonormal basis for V_x. These eigenvectors are linear combinations of the variables x_i, and are called PC directions. The eigenvector corresponding to the largest eigenvalue is the direction in V_x of maximum variability, and so on, in decreasing order of the magnitudes of the eigenvalues. The transpose of the m eigenvectors corresponding to the zero eigenvalues represents a basis for $V_x^\perp$. Note that these eigenvectors are not uniquely defined, because the corresponding eigenvalues are all equal. Although in some applications the PC directions may be useful, from the viewpoint of model identification they do not have any advantage over any

[2]The term error covariance matrix is used throughout the text instead of the commonly used term variance–covariance matrix.

other basis choice. Nevertheless, PCA does identify a basis for $V_x^{\perp}$ exactly in the absence of measurement errors.

Once a basis for V_x is obtained, the regression matrix can also be easily obtained for any specified set of $n-m$ independent variables as follows. Without loss of generality, let the last $n-m$ variables be specified as the independent variables. Let U_r be the basis matrix identified for V_x (whose columns are the orthonormal eigenvectors corresponding to the $n-m$ nonzero eigenvalues). The matrix U_r can be partitioned as

$$U_r = \begin{bmatrix} U_D \\ U_I \end{bmatrix} \begin{matrix} m \\ n-m. \end{matrix} \tag{6}$$

The regression matrix is then given by

$$B = U_D(U_I)^{-1}. \tag{7}$$

If U_I is singular, then this implies that the independent set of variables is specified incorrectly. In other words, there exist linear relationships between the specified 'independent' set of variables. An important feature of using PCA for multivariate regression is that the independent and dependent variables are treated alike and no distinction is made between them (also referred to as symmetric treatment of variables). Furthermore, if there are linear dependencies between the specified 'independent' variables, either due to incorrect specification or due to the characteristics of the data set obtained, then PCA identifies such relationships also, and is therefore a more general method for multivariate regression.

3. Effect of scaling in PCA

If the data are scaled before applying PCA, then the question can be raised as to whether it is possible to obtain an exact basis for $V_x^{\perp}$ in the absence of measurement errors. In order to answer this question, the following general linear transformation of the data is considered:

$$y_s(t) = Dy(t) = Dx(t) = x_s(t), \tag{8}$$

where D is any nonsingular matrix. If D is diagonal, then the above transformation defines a scaling of the data. PCA can be applied to the covariance matrix S_{y_s}, which is defined as

$$S_{y_s} = \frac{1}{N}\sum_{t=1}^{N} y_s(t)y_s^{\mathrm{T}}(t) = \frac{1}{N} Y_s Y_s^{\mathrm{T}}, \tag{9}$$

where the scaled data matrix Y_s is defined in a manner analogous to Eq. (5). Since D is nonsingular, the rank of S_{y_s} is also equal to $n-m$. Thus, by applying PCA to S_{y_s}, a basis for the space orthogonal to the scaled data vector $x_s(t)$ is obtained using the transpose of the m orthonormal eigenvectors corresponding to the zero eigenvalues. If the transpose of these eigenvectors is denoted by A_s, then

$$A_s x_s(t) = 0. \tag{10}$$

Using Eq. (8) in the above equation, the following relation is obtained:

$$A_s D x(t) = 0. \tag{11}$$

From the above equation, it can be deduced that the rows of the matrix $A = A_s D$ are a basis for $V_x^{\perp}$. Thus, in the absence of measurement errors, an exact basis for $V_x^{\perp}$ is obtained even if PCA is applied to transformed (or scaled) data using Eq. (8). However, it must be noted that the rows of A are not orthonormal and they also do not correspond to the eigenvectors of S_y.

4. Model identification with known Σ_ε

The problem of model identification from noisy measurements using PCA is now considered, under the assumption that the measurement error covariance matrix, Σ_ε, is known. If measurements are noisy, then S_y will be a full rank matrix, and by using PCA it is not possible to obtain an exact basis for V_x or $V_x^{\perp}$. In fact, it is not possible to establish a relationship between the orthonormal eigenvectors of S_y and those of S_x. Furthermore, if the data are transformed or scaled using Eq. (8), the eigenvectors of S_{y_s} and those of S_y do not bear any simple relation to each other (Morrison, 1967). Both these problems have been hitherto tackled in a heuristic manner in model identification from noisy data using PCA. If it is assumed that the error variances are much smaller compared to the variances in $x(t)$, then S_y will possess $n-m$ dominant eigenvalues and m small eigenvalues. The orthonormal eigenvectors corresponding to the small eigenvalues can be used as an estimate for the basis of $V_x^{\perp}$. It has also been suggested that if x contains variables which are not commensurate, then it is better to scale the data using standard deviations of the measurements. Other scaling strategies have also been suggested which can be applied under restrictive assumptions (Wentzell et al., 1997). The effect of these heuristics on the quality of the identified model cannot be easily assessed. In what follows, a procedure is described which effectively resolves the issue of appropriately scaling noisy data, such that a basis for $V_x^{\perp}$ can be exactly obtained using PCA, under the assumption that Σ_ε is known.

Let L be the Cholesky factor of Σ_ε defined by

$$LL^{\mathrm{T}} = \Sigma_\varepsilon. \tag{12}$$

Similar to Eq. (8), the measurements are transformed using L^{-1} as the nonsingular transformation matrix. The transformed measurements are given by

$$\begin{aligned} y_s(t) &= L^{-1}y(t) = L^{-1}x(t) + L^{-1}\varepsilon(t) \\ &= x_s(t) + L^{-1}\varepsilon(t). \end{aligned} \tag{13}$$

If Σ_ε is a diagonal matrix, then L is also a diagonal matrix containing the standard deviations of measurement errors, and the above transformation is equivalent to scaling the data using standard deviations of the corresponding measurement errors.

By taking the expectation of S_{y_s}, it can be easily shown that

$$\Sigma_{y_s} = S_{x_s} + I. \tag{14}$$

In the above equation Σ_{y_s} is the population covariance matrix of y_s, while $S_{x_s} = L^{-1} S_x L^{-\mathrm{T}}$. It may be noted that the expectation of S_{x_s} is the same as S_{x_s}, since $x(t)$ is assumed to be a deterministic sequence.

From Eq. (14) and the eigenvalue shift theorem, the following two important results can be immediately derived:

1. The eigenvectors of Σ_{y_s} are identical to those of S_{x_s}.
2. The eigenvalues of Σ_{y_s} are equal to the corresponding eigenvalues of S_{x_s} increased by unity.

Since S_{x_s} is of rank $n-m$ it will have m zero eigenvalues. From the above results, it can be concluded that the corresponding eigenvalues of Σ_{y_s} will be unity. Furthermore, the eigenvectors, corresponding to the eigenvalues of Σ_{y_s} that are greater than unity, define a basis for V_{x_s}. It has been shown already that a basis for V_x can be obtained exactly, given the basis for V_{x_s}. Thus, given a sample of measurements, the measurements can be transformed as in Eq. (13) and PCA applied to S_{y_s}. The eigenvectors corresponding to the eigenvalues that are close to unity can be used to obtain a basis for $V_x^{\perp}$ (refer to the discussion that follows Eq. (11)). Using Theorem 2.3 (Ljung, 1999), it can be proved that S_{y_s} is a consistent estimate of Σ_{y_s}. Thus, in the limit as the sample size goes to infinity, an exact basis for $V_x^{\perp}$ is obtained using this method. The above scaling procedure and application of PCA for multivariate regression is well known in statistical literature (Fuller, 1987, Theorem 4.1.1), and gives maximum likelihood estimates of the model parameters.

Wentzell et al. (1997) proposed an MLE technique for model identification using PCA when the covariance matrix of measurement errors is known, and the model order is also specified. Their procedure is an alternating regression procedure which does not scale the data. Instead, their method iteratively transforms the model identified by applying PCA to unscaled data, until the maximum likelihood estimates of $x(t)$ are obtained. In contrast, the procedure that is described above is a non-iterative technique, which has a stronger theoretical basis and also provides additional useful information. In particular, the fact that the eigenvalues of S_{y_s} corresponding to the eigenvectors which define a basis for $V_x^{\perp}$ should be unity can be used to obtain the model order m. If an incorrect value of m is assumed, then the eigenvalues corresponding to the last m eigenvectors of S_{y_s} may not be close to unity. It should, however, be pointed out that the method proposed by Wentzell et al. (1997) is applicable to general measurement error covariance structures such as the cases when the error variances are different for different variables, change from sample to sample, and are correlated across samples and variables, as long as the variances and covariances are completely specified. On the other hand, the method based on scaling strategy proposed here is applicable only if the covariance matrix is constant along one direction (either with respect to samples or with respect to variables). The advantage of the proposed scaling procedure becomes evident in the next section where it is extended to simultaneously estimate the model and unknown measurement error covariance matrix from data.

5. Simultaneous model identification and error covariance matrix estimation

If Σ_ε is unknown, then the method described in the preceding section can be applied, if the error covariance matrix can be estimated from the data along with the model. However, it has been proved (Fuller, 1987, p. 104) that even in the case of univariate regression for the functional model problem, the maximum likelihood procedure for simultaneously estimating the model and the error variances gives an unbounded solution. Thus, a modification of the MLE procedure is required to simultaneously estimate both the model and error covariance matrix (Chan & Mak, 1983). The procedure proposed here is based on constructing the likelihood function of the constraint residuals, and determining the estimates of the model and error covariance matrix which maximizes this likelihood function. Based on the insights of the preceding section, an iterative algorithm is proposed which combines PCA with an optimization procedure for estimating the error covariance matrix that maximizes the likelihood function of the constraint residuals. It may be noted that when the error covariance matrix is known, then the model can be estimated using a non-iterative method, by applying PCA to scaled residuals as described in the preceding section. However, since the model and the error covariance matrix are both unknown, an iterative algorithm is proposed to estimate them simultaneously. A description of the proposed algorithm follows.

It is assumed that an initial estimate of the model constraint matrix, $\hat{A}^0$, is available. For example, an initial estimate can be obtained by applying PCA to the measured data. Using this initial model estimate, the constraint residuals are computed at each time instant as

$$r(t) = \hat{A}^0 y(t). \tag{15}$$

If the estimated model is exact, then the constraint residuals will be independent normally distributed variables with zero mean and covariance matrix $\Sigma_r = \hat{A}^0 \Sigma_\varepsilon (\hat{A}^0)^{\mathrm{T}}$. Thus, the joint density function of $r(1) \ldots r(N)$ can be easily obtained (Morrison, 1967), and an estimate of Σ_ε can be obtained by maximizing the log likelihood function of $r(1) \ldots r(N)$. This results in the following nonlinear optimization problem:

$$\min_{\Sigma_\varepsilon} N \log |\hat{A}^0 \Sigma_\varepsilon (\hat{A}^0)^{\mathrm{T}}| + \sum_{t=1}^{N} (r^{\mathrm{T}}(t) (\hat{A}^0 \Sigma_\varepsilon (\hat{A}^0)^{\mathrm{T}})^{-1} r(t)). \tag{16}$$

The above MLE problem can also be interpreted as a procedure for extracting an estimate of Σ_ε, given an estimate of the covariance matrix of constraint residuals Σ_r. This follows from the fact that the maximum likelihood estimate of Σ_r (which maximizes the likelihood function of $r(1)\ldots r(N)$) is the sample covariance matrix S_r. The estimate of Σ_ε, which maximizes the same likelihood function, is the one that satisfies the following relation:

$$\hat{A}^0\hat{\Sigma}_\varepsilon(\hat{A}^0)^{\mathrm{T}} = S_r. \tag{17}$$

Depending on the number of constraints and number of variables, it may or may not be possible to satisfy the above equation. The number of independent equations involved in Eq. (17) is $m(m+1)/2$, due to both S_r and Σ_ε being symmetric matrices. Therefore, these equations can be exactly satisfied only if the number of unknown parameters in Σ_ε that need to be estimated is exactly equal to $m(m+1)/2$. Typically, the number of constraints m is less than the number of variables n. Thus, if estimates of all diagonal and off-diagonal elements of Σ_ε are required, then multiple solutions that satisfy Eq. (17) are obtained. A reasonable assumption that can be made is that the errors in different variables are independent. In this case, Σ_ε is a diagonal matrix, and the number of unknown parameters to be estimated is n. Even in this case, a non-degenerate estimate for Σ_ε is obtained only if $m(m+1)\geqslant 2n$. Other techniques, based on least squares criterion, have been proposed for estimating the measurement error covariance matrix, given the constraint model and the covariance matrix of constraint residuals (Romagnoli & Sanchez, 1999). However, these methods are not maximum likelihood estimates.

Assuming that the number of diagonal and off-diagonal elements of Σ_ε that have to be estimated is less than or equal to $m(m+1)/2$, function (16) can be minimized. Lower bounds can also be imposed on the elements of Σ_ε that are estimated, and the constrained optimization problem can be solved. Let the estimate of measurement error covariance matrix obtained using the above method be denoted as $\hat{\Sigma}_\varepsilon^0$. It should be noted that this estimate has been obtained assuming that the model has been estimated exactly. The estimated matrix $\hat{\Sigma}_\varepsilon^0$ can be used to transform the data as described in the preceding section, and PCA applied to the transformed measurements in order to obtain an updated estimate of the constraint matrix. The entire procedure can be repeated until the estimates for the model and error covariance matrix 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.

The proposed iterative algorithmic procedure can be summarized as follows. The notation $[U,S,V]=\mathrm{svd}(Y)$ is used to denote the singular value decomposition of an $n\times N$ data matrix Y, where U is the $n\times n$ matrix of left singular vectors, S is an $n\times n$ matrix containing the n nonzero singular values ordered from the largest to the smallest along the diagonal, and V is an $n\times N$ matrix of right singular vectors.

Step 1: Set iteration counter $k=1$ and λ^0 to be zero.

Step 2: Set estimates of the nonzero elements of Σ_ε^k to be a small fraction (say 0.0001) of the corresponding elements of S_y.

Step 3: Obtain transformed matrix $Y_s = L^{-1}Y$ where $LL^{\mathrm{T}} = \Sigma_\varepsilon^k$.

Step 4: Let $[U,S,V]=\mathrm{svd}(Y_s)$. Obtain estimate $A^k = U_{n-m+1\ldots n}^{\mathrm{T}}L^{-1}$, where $U_{n-m+1\ldots n}$ is the sub-matrix of U corresponding to the last m columns.

Step 5: Let λ^k be the sum of the last m singular values. Stop if relative change in λ is less than specified tolerance; else continue.

Step 6: Obtain the solution for the nonzero elements of Σ_ε by minimizing the function $N\log|A^k\Sigma_\varepsilon(A^k)^{\mathrm{T}}|+\sum_{t=1}^{N} r(t)^{\mathrm{T}}(A^k\Sigma_\varepsilon(A^k)^{\mathrm{T}})^{-1}r(t)$. Denote the solution as Σ_ε^{k+1}.

Step 7: Increment iteration counter k and return to Step 3.

In the optimization Step 6, if positivity constraints are imposed on the estimates of the nonzero elements of Σ_ε and that they should also satisfy $A^k\Sigma_\varepsilon(A^k)^{\mathrm{T}}=I$, then it can be proved that the above iterative procedure estimates the model and error covariance matrix simultaneously by maximizing the constrained likelihood function of the constraint residuals. The proof, which is described in the Appendix, also proves that the iterative algorithm converges. It should be noted that, in general, it may be possible to obtain only a local optimum of the optimization problem in Step 6, and not a global optimum.

It should be noted that in order to apply the above algorithm, the model order has to be specified (since even an initial guess of the constraint matrix using PCA can be obtained only if the model order is given). In order to determine the true model order, the above algorithm can be applied for different guesses of the model order. For each guess of the model order $\hat{m}$, the converged singular values can be examined to verify whether $\hat{m}$ unity singular values are obtained. The application of this procedure to determine the unknown model order and corresponding model and error covariance matrix is illustrated by means of an example in the following section.

6. Simulation results and discussion

Example 1. The first example that is chosen is a simple flow process example shown in Fig. 1. The above example has been chosen so that it satisfies the condition $m(m+1)>2n$ (in the above example $m=3$ and $n=5$). It is assumed that the measurement error covariance matrix is diagonal. The constraint matrix for this process is given by

$$A=\begin{bmatrix} 1 & 1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1 & 0\\ 0 & -1 & 0 & 1 & -1 \end{bmatrix}. \tag{18}$$

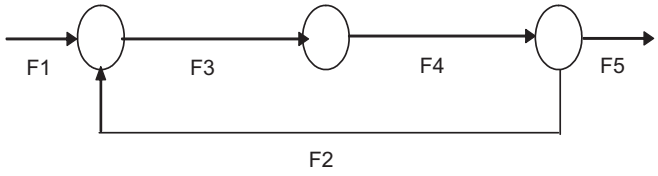

Fig. 1. Schematic of a flow process.

In order to simulate the true values of variables at each time instant, a set of independent flow variables are chosen (in the above example F1 and F2 are chosen as independent variables). The true values of independent variables are simulated by adding normally distributed random fluctuations to their base values. The true values of the dependent flow variables are calculated such that they satisfy the flow balance constraints. The base values of variables and the standard deviations of the fluctuations are given in Table 1. In the simulation procedure, the measured values of variables are simulated by adding normally distributed random noise to their true values. The standard deviations of measurement errors are also given in Table 1. A sample of 1000 measurement vectors is simulated and the procedure described in Section 5 is applied.

In order to evaluate the accuracy of the estimated basis for $V_x^{\perp}$, the distance between the row spaces of the true constraint matrix and the estimated constraint matrix can be used. The minimum distance of each row of A from the subspace spanned by the rows of $\hat{A}$ is given by

$$\alpha_i = \|A_{i.} - A_{i.}\hat{A}^{\mathrm{T}}(\hat{A}\hat{A}^{\mathrm{T}})^{-1}\hat{A}\|. \tag{19}$$

A consolidated measure of model estimation accuracy is given by

$$\alpha = \sum_{i=1}^{m} \alpha_i. \tag{20}$$

The above measure treats all bases sets for the row space of $\hat{A}$ as equivalent, although it does depend on the basis choice for the row space of true constraint matrix A. Alternatively, the angle θ between the row spaces of A and $\hat{A}$ can also be used as a measure of the model estimation accuracy. This measure is invariant with respect to both the bases choice in the row space of the true constraint matrix and the row space of the estimated constraint matrix.

The results obtained for the above example using PCA for different choices of data scaling are presented in Table 2. Results obtained using the proposed iterative method (denoted as IPCA) for the same example are also presented in this table. In both approaches, the actual number of constraints are assumed to be known.

In Table 2, the first three rows are the results obtained using PCA, respectively, when the measured data are not scaled, scaled using sample standard deviation of the corresponding measurement, and scaled using true standard deviations of measurement errors. The last row gives the results obtained using the proposed method. The constraint matrix obtained by PCA is used as an initial estimate in IPCA. From the values of α and θ, it can be concluded that a good estimate of the model constraints is obtained using both PCA and IPCA. This is due to the fact that in this simulation the signal-to-noise variation is high (ratio of their standard deviations is more than 10). However, even in this case, the proposed iterative method is able to improve the accuracy of the model (in terms of the angle θ) obtained through PCA by more than 80%. The number of major iterations required for IPCA to converge was around 10, although within three to four iterations the estimates obtained are very close to the final converged values. The estimated standard deviations of measurement error variances obtained using the proposed method are [0.1121 0.0837 0.1406 0.2031 0.1775], which are close to their true values. For the given sample of data, the best achievable model accuracy is obtained when the data are scaled using the true standard deviations of measurement errors, as shown in the third row of Table 2. It is observed that the accuracy of model obtained using IPCA is very close to this achievable limit. The same simulations were also repeated with sample sizes larger than 1000. Although these results are not presented, it was observed that as the sample size increased, the estimates of both the constraint matrix and error covariance matrix converged to their true values indicating that statistically consistent estimates are obtained. However, the statistical consistency of the estimates of the model and error covariance matrix has to be proved theoretically.

The converged singular values, [236.5 17.7 1.01 1.0 0.99], obtained using IPCA reveal an interesting feature. It can be observed that the singular values corresponding to the last three PCs (which correspond to the assumed number of constraints) are very close to unity, as theoretically predicted. In contrast, the singular values

Table 1
Data for simulating true values of variables

Flow variable	True values		σ_ε
	Base value	Std. of fluctuation	
F1	10	1.0	0.1
F2	10	2.0	0.08
F3		F1 + F2	0.15
F4		F3	0.2
F5		F4 − F2	0.18

Table 2
Quality of the model identified for different scaling choices

Case	Scale	$\alpha \times 10^3$	θ (deg)
PCA	None	5.86	0.17
PCA	σ_y	10.22	0.24
PCA	σ_ε	1.62	0.028
IPCA		1.2	0.03

obtained using PCA for the three scaling strategies are, respectively, [33.32 1.9 0.18 0.16 0.11], [19.84 1.35 0.15 0.08 0.06], and [238.9 18.8 1.06 0.99 0.97]. Clearly, by scaling the data differently, the singular values can be altered, and this may make it difficult to determine the number of PCs to be retained/rejected. In other words, it may not be possible to determine the number of constraints precisely by examining the singular values of the scaled data, unless the standard deviations of measurement errors are used for scaling. It may also be noted that if the data are auto-scaled, a worse model may be obtained compared to the case when the data are not scaled at all (compare results of first and second rows of Table 2).

In order to evaluate how the proposed method performs for low signal-to-noise variation, the standard deviations of the true value variations in F1 and F2 are reduced to 0.2 each, while retaining the standard deviations of measurement errors as before. The results obtained for this case are given in Table 3. As expected the accuracy of the models estimated by both approaches has decreased. However, a good estimate of the model is still obtained using the proposed approach, and there is a 90% improvement over the model obtained using PCA. The estimated standard deviations of measurement errors using the proposed approach are [0.1121 0.0838 0.1406 0.2031 0.1774], which are same as before. Thus, even though the model is estimated less accurately, the measurement error standard deviations are estimated fairly accurately by the proposed approach. The converged singular values obtained are [233.5 2.4 1.01 1.0 0.98], which again satisfy the condition that the singular values corresponding to the assumed number of constraints are close to unity.

In order to demonstrate that the proposed method can be used even if errors in different variables are correlated, data were simulated for the above example using an error covariance structure which contained an off-diagonal element. It should be noted that since the above process has only three constraints, at most six elements of the error covariance matrix can be estimated. This implies that besides the diagonal elements, at most one off-diagonal element can be estimated from the measured data. The true flow rates in this case are simulated as described in Table 1. The nonzero elements of the measurement error covariance are chosen as [0.0244 0.0064 0.0369 0.04 0.0324 0.03], where the first five elements are the diagonal elements (error variances) and the last element is the covariance between errors in variables 1 and 3. The results for this case are shown in Table 4.

The above results again indicate that the model obtained using IPCA is better than that obtained using PCA, and is close to the maximum achievable accuracy. The nonzero elements of the estimated error covariance matrix are [0.0266 0.007 0.0367 0.0402 0.0312 0.0312], which are also close to their corresponding true values. The converged singular values obtained using IPCA are [2458.5 27.53 1 1 1]. As theoretically predicted, the last three singular values are unity even in this case.

It was stated in the preceding section that it is also possible to determine the number of constraints using the proposed approach. As a test of this, the number of constraints was incorrectly assumed as four instead of three in the above simulations, and the proposed method was used. In this case, the estimated standard deviations of measurement errors obtained are [1.118 1.117 0.047 0.242 1.127], and the singular values are [434.37 1.72 1.00 0.15 0.11]. Since the singular values corresponding to the last four eigenvalues are not close to unity, this indicates that the number of constraints has been incorrectly assumed. An outer iteration to estimate the model order correctly is demonstrated using a larger example below.

Example 2. As a second example a much larger flow process as shown in Fig. 2 is chosen. This process is a steam metering network of a methanol synthesis plant and has been used as a popular case study for testing steady-state data reconciliation and gross error detection strategies (Serth & Heenan, 1986). It consists of 28 flows related by 11 flow balance constraints. It is assumed that all the flows are measured. The true flow rates of streams F4, F6, F10, F11, F13, F14, F16–F22, F24, F26–F28 are chosen as the independent variables and simulated as a first order AR sequence given by

$$x_{i,k+1} = a_i x_{i,k} + b_i w_{i,k}, \tag{21}$$

where $w_{i,k}$ is an independent standard normal random variable. The remaining 11 flow rates are calculated at each time using the flow balance equations. It should be noted that the true values simulated in this manner are autocorrelated. The measured values are simulated by adding independent normally distributed random errors with mean zero and standard deviation ε_i to the true values. Table 5 gives the parameter values used in the simulation. The standard deviation of the measurement error in a variable is chosen to be 2.5% of the corresponding true

Table 3
Quality of the model identified for low signal-to-noise ratio

Case	Scale	$\alpha \times 10^3$	θ (deg)
PCA	None	447.0	12.73
PCA	σ_y	252.1	7.61
PCA	σ_ε	21.1	0.49
IPCA		32.5	1.39

Table 4
Model identification for non-diagonal error covariance matrix

Case	Scale	$\alpha \times 10^3$	θ (deg)
PCA	None	10.53	0.20
PCA	σ_y	13.68	0.27
PCA	Cholesky factor of Σ_ε	1.21	0.024
IPCA		1.47	0.043

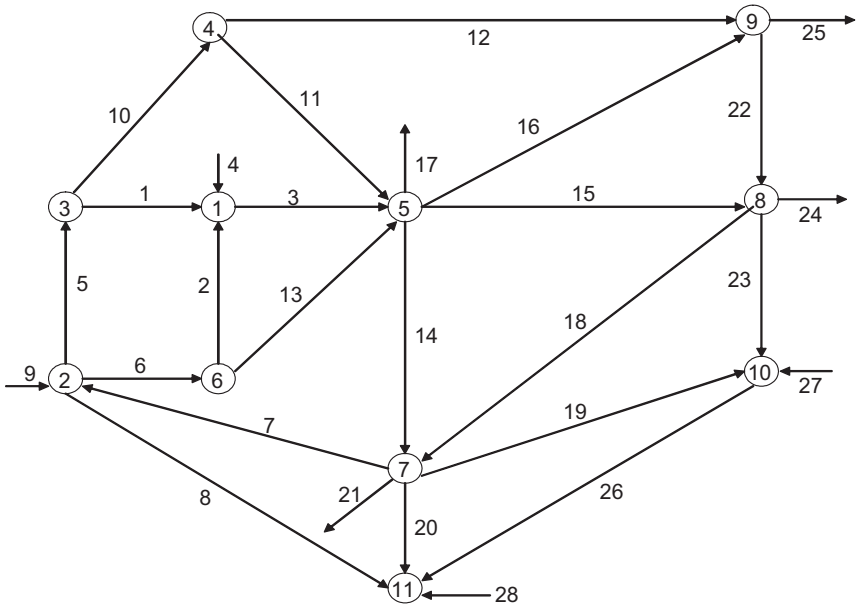

Fig. 2. Steam metering network of methanol synthesis plant.

Table 5
Parameter values used for simulating steam metering network

Flow variable	Initial true value	a_i	b_i
F4	109.95	0.9975	3.4497
F6	112.27	0.9616	2.7893
F10	52.41	0.9803	1.1375
F11	14.86	0.9743	7.7211
F13	111.27	0.9946	5.0059
F14	91.86	0.9881	9.3863
F16	23.64	0.9728	5.1939
F17	32.73	0.9509	4.7678
F18	16.23	0.9911	8.6160
F19	7.95	0.9722	5.7264
F20	10.5	0.9808	2.8238
F21	87.27	0.9896	7.0492
F22	5.45	0.9961	8.5431
F24	46.64	0.9869	1.1768
F26	81.32	0.9588	7.1315
F27	70.77	0.9703	4.4153
F28	72.23	0.9968	8.4862

initial flow rate. A total of 1000 samples are generated. The signal-to-noise ratios (defined as the ratio $\sigma_{x_i}/\sigma_{\varepsilon_i}$) of the different variables span a range 6–7550 for the sample generated. For this process, the proposed IPCA algorithm is applied for different guesses of the model order, and the model is estimated along with the elements of the error covariance matrix. Assuming the error covariance matrix to be diagonal, $n = 28$ error variances have to be estimated. For estimating these variances, along with the model at least $m = 7$ constraints are required, since this will ensure that the inequality $m(m+1) \geqslant n$ is satisfied. Thus, the minimum model order that can be assumed is 7. It may be noted that the true order for this process is 11. Fig. 3 shows a plot of the logarithm of the converged singular values

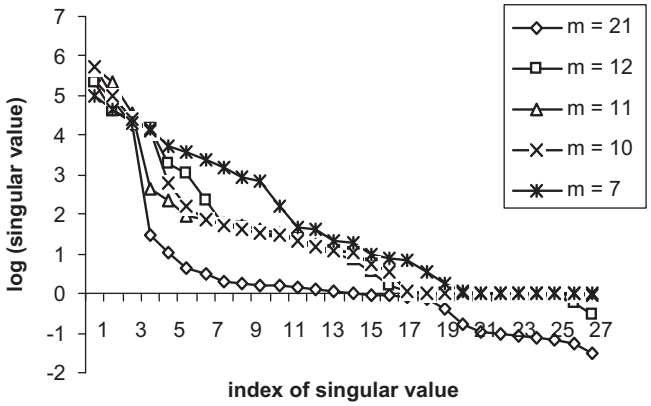

Fig. 3. Singular values obtained using IPCA for different model orders.

obtained using IPCA for different guesses of the model order. From Fig. 3 it can be observed that if the model order guessed is greater than the true order, then the number of unity singular values obtained after convergence is not equal to the model order guessed. For example, if the guess of model order is 21, then many of the smaller singular values are much less than unity. Even if the model order guessed is one more than the true order ($m = 12$), then the last two singular values obtained are equal to 0.56 and 0.3, which are both less than unity. However, if the model order guessed is equal to or less than the true order, then as many unity singular values are obtained as the model order guessed. Therefore, a necessary (but not sufficient) condition for the guessed model order to be correct is that the number of unity singular values obtained at convergence of the proposed algorithm be equal to the guessed model order. Violation of this condition implies that the true order has not been correctly estimated.

A systematic procedure can now be devised for determining the true process order using one of the following methods:

- Start with the smallest possible model order necessary for estimating the measurement error variances and successively increment it, until the number of unity singular values obtained using IPCA algorithm exceeds the guess of the model order.
- Alternatively, start with a high estimate of the model order and successively decrement it until the condition that the number of unity singular values obtained using IPCA equals the guess of the model order is satisfied.

The above approaches can also be suitably combined to obtain a more efficient algorithm for determining the true process order.

7. Concluding remarks

In this paper, an algorithm is proposed for simultaneously estimating an accurate process model and the measurement error covariance matrix from noisy data, using an iterative PCA technique. As part of the development, the outstanding issue of appropriately scaling or transforming noisy data before applying PCA has also been resolved. A new criteria for determining model order by examining the singular values obtained using PCA on the transformed data is proposed, which has a rigorous theoretical basis.

The above important advantages of the proposed method are obtained provided the following three conditions are satisfied:

- The underlying relationships that relate the variables are linear.
- The measurement errors from one sample to another are mutually independent.
- The measured samples correspond to different steady states, which implies that the true values corresponding to each sample strictly satisfy Eq. (1).

In practice, one may wish to obtain an approximate linear model of a mildly nonlinear process. Moreover, processes are never truly operating at a steady state and usually fluctuate around a nominal steady state. In such cases, the covariance matrix of constraint residuals will include the effect of measurement errors as well as the effect of process variability. In such cases, it is necessary to investigate whether an estimate of the measurement error covariance matrix can be extracted from the constraint residuals, or alternatively investigate the robustness of the proposed approach to such deviations and determine whether the theoretical advantages of the proposed approach can be realized in practice.

Acknowledgements

Financial support from NSERC, Matrikon Inc., and ASRA in the form of the Industrial Research Chair Program at the University of Alberta is gratefully acknowledged.

Appendix

In this Appendix, a proof is presented that shows that the proposed IPCA method maximizes the constrained likelihood function of the residuals to obtain the estimates of the model and error covariance matrix. The proof also establishes the fact that the iterative algorithm used in IPCA converges.

Given a set of measurements $y(1), y(2), \ldots, y(N)$ which satisfy Eqs. (1) and (2), the joint likelihood function of the constraint residuals can be written as

$$-2\log L(r(1), r(2), \ldots, r(N)) = N\log|A\Sigma_\varepsilon A^{\mathrm{T}}| + \sum_{t=1}^{N} r(t)^{\mathrm{T}}(A\Sigma_\varepsilon A^{\mathrm{T}})^{-1}r(t), \quad (22)$$

where $r(t)$ are the constraint residuals defined in Eq. (15). The estimates of A (or the m dimensional subspace of the R^n for which the rows of A form a basis) and Σ_ε which minimize the above likelihood function have to be obtained. This optimization problem can be solved iteratively, by first considering Σ_ε to be known and solving for A.

For a given Σ_ε, the likelihood function given by Eq. (22) can be written in terms of transformed variables as

$$-2\log L(r(1), r(2), \ldots, r(N)) = N\log|A_s A_s^{\mathrm{T}}| + \sum_{t=1}^{N} y_s(t)^{\mathrm{T}} A_s^{\mathrm{T}}(A_s A_s^{\mathrm{T}})^{-1}A_s y_s(t), \quad (23)$$

where y_s is transformed as defined by Eq. (13) using the Cholesky factor L of Σ_ε, and $A_s = AL$. An estimate of A_s (or a basis for the m dimensional row space of A_s) can be obtained by minimizing Eq. (23). Without loss of generality, the rows of A_s can be chosen to be orthonormal, which implies that the first term in the RHS of Eq. (23) is a constant and can be disregarded. The rows of A_s which minimize the second term can be obtained by choosing them to be the orthonormal eigenvectors of $S_{y_s} = (1/N)Y_s^{\mathrm{T}}Y_s$ corresponding to the last m smallest eigenvalues (Rao, 1964). It should be noted that this minimization is performed in *Step* 4 of the algorithm.

In the second stage Eq. (22) can be minimized, keeping the value of A fixed at the optimum value obtained in previous step, to obtain an updated estimate of Σ_ε. In the algorithm, this is done in *Step* 6. In this optimization step, if the constraint that $A_s A_s^{\mathrm{T}} = I$ for the given estimate of A is also imposed, then this implies that the same function is being minimized at each step starting with the current optimal values of the parameters. This ensures that the

function value is non-increasing at each step and therefore the algorithm must converge. Furthermore, the converged solution maximizes the constrained likelihood function of the residuals, since this is the function being minimized in both steps.

References

Brauwere, A. D., Pintelon, R., Ridder, F. D., Schoukens, J., & Baeyens, W. (2007). Estimation of heteroscedastic measurement noise variances. *Chemometrics and Intelligent Laboratory Systems*, *86*, 130–138.

Chan, N. N., & Mak, T. K. (1983). Estimation of multivariate linear functional relationships. *Biometrika*, *70*(1), 263–267.

Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.

Liu, J. (2007). On-line soft sensor for polyethylene process with multiple production grades. *Control Engineering Practice*, doi:10.1016/j.conengpractice.2005.12.005.

Ljung, L. (1999). *System identification: Theory for the user* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Morrison, D. F. (1967). *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill.

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

Romagnoli, J. A., & Sanchez, M. C. (1999). *Data processing and reconciliation for chemical process operation*. New York: Academic Press.

Serth, R. W., & Heenan, W. A. (1986). Gross error detection and reconciliation in steam-metering systems. *AIChE Journal*, *32*, 733.

Viberg, M. (1995). Subspace-based method for the identification of linear time-invariant systems. *Automatica*, *31*(12), 1835–1851.

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

Yoon, S., & MacGregor, J. F. (2000). Statistical and causal model-based approaches to fault detection and isolation. *AIChE Journal*, *46*(9), 1813–1899.