MODEL IDENTIFICATION AND ERROR COVARIANCE MATRIX ESTIMATION FROM NOISY DATA USING PCA

Shankar Narasimhan *,[1] and **Sirish L. Shah** *

* *Department of Chemical & Materials Engineering, University of Alberta, Edmonton, AB, CANADA*

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, we develop the theoretical basis and an iterative algorithm for model identification using PCA, 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 is a multivariate statistical tool developed primarily to obtain a parsimonious representation of multivariate data. This is achieved by choosing a few linear combinations known as principal components, 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. 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 and 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 *et al.*, 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 are likely to change over time.

In this paper, we describe an iterative method which combines PCA with a maximum likelihood estimation procedure for obtaining an estimate

[1] Visiting Professor from Department of Chemical Engineering, IIT Madras, Chennai, INDIA

of the error covariance matrix. The proposed approach also provides answers to 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

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

Let $x(t)$ be a set 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

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

We assume that the random errors, $\epsilon(t)$, are temporally independent and follow a multivariate normal distribution with mean zero and covariance matrix Σ_ϵ. The random errors are also assumed to be independent of $x(t)$. Given a sample of N measurements, $y(1) \dots y(N)$, the objective is to estimate the constraint matrix (also referred to as the model).

We assume that the true values of variables, $x(t)$, are a deterministic sequence satisfying the following two conditions.

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

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

where $\bar{x}$ represents the average of the sequence $x(t)$, and μ_x and Σ_x are bounded. The above assumptions ensure that $y(t)$ is a quasi-stationary signal (Ljung, 1999).

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 a m dimensional subspace of R^n (denoted as V_c), 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_c. It can be further noted that in order to define V_x and V_c, we only need to identify a basis for each of these spaces. 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 we have a sample of $n-m$ linearly independent realizations of $x(t)$, then we can use it as a basis for V_x. We can then construct m linearly independent vectors orthogonal to V_x, which define a basis for V_c exactly. Note that this is sufficient to solve the stated problem.

If we use PCA to solve the above problem, then we will determine the orthonormal eigenvectors of the data variance matrix $S_y = \frac{1}{N} Y^T Y$ (which is identical to $S_x = \frac{1}{N} X^T X$ in the absence of measurement errors), where

$$Y = [y(1), y(2), \dots, y(N)]^T \tag{3}$$

Since the column space of S_y is identical to V_x, 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 principal component 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 represent a basis for V_c. 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 other basis choice. Nevertheless, PCA does identify a basis for V_c exactly in the absence of measurement errors.

3. EFFECT OF SCALING IN PCA

We can raise the question of whether we can obtain an exact basis for V_c in the absence of measurement errors, if we scale the data before applying PCA. In order to answer this question, we will consider the following general linear transformation of the data

$$y_s(t) = Dy(t) = Dx(t) = x_s(t) \tag{4}$$

where D is any nonsingular matrix. If D is diagonal, then the above transformation defines a

scaling of the data. We can apply PCA to the variance matrix $S_{y_s} = \frac{1}{N} Y_s^T Y_s$ where the scaled data matrix Y_s is defined in a manner analogous to eq. 3. Since D is nonsingular, the rank of S_{y_s} is also equal to $n-m$. Thus, if we apply PCA using S_{y_s}, the transpose of the m orthonormal eigenvectors corresponding to the zero eigenvalues represent a basis for the space orthogonal to the scaled data vector $x_s(t)$. If we denote the transpose of these eigenvectors by A_s, then we can write

$$A_s x_s(t) = 0 \tag{5}$$

Using eq. 4 in the above equation we get

$$A_s D x(t) = 0 \tag{6}$$

From the above equation, we can deduce that the rows of the matrix $A = A_s D$ is a basis for V_c. Thus, in the absence of measurement errors, we obtain an exact basis for V_c even if we apply PCA to transformed (or scaled) data using eq. 4. 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

We now consider the problem of model identification from noisy measurements using PCA, 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 we will not be able to obtain an exact basis for V_x or V_c. In fact, it is not possible to establish a relationship between the orthonormal eigenvectors of S_y and those of S_x. Furthermore, if we scale or transform the data using eq. 4, 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 we assume that the error variances are much smaller compared to the variances in $x(t)$, then we can expect S_y to 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_c. 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, we describe a procedure which effectively resolves the issue of appropriately scaling noisy data, such that a basis for V_c can be exactly obtained using PCA, under the assumption that Σ_ϵ is known.

Let L be the square root of Σ_ϵ defined by

$$LL^T = \Sigma_\epsilon \tag{7}$$

Similar to eq. 4, we will transform the measurements using L^{-1} as the nonsingular transformation matrix. The transformed measurements are given by

$$\begin{aligned} y_s(t) = L^{-1} y(t) &= L^{-1} x(t) + L^{-1} \epsilon(t) \\ &= x_s(t) + L^{-1} \epsilon(t) \end{aligned} \tag{8}$$

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

$$\Sigma_{y_s} = S_{x_s} + I \tag{9}$$

In the above equation Σ_{y_s} is the population variance matrix of y_s, while $S_{x_s} = L^{-1} S_x L^{-T}$ (since $x(t)$ is deterministic).

From eq. 9 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, we can conclude 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}. We have already shown that we can obtain the basis for V_x exactly, given the basis for V_{x_s}. Thus, given a sample of measurements, we can transform the measurements as in eq. 8 and apply PCA on S_{y_s}. The eigenvectors corresponding to the eigenvalues that are close to unity, can be used to obtain a basis for V_c (refer to the discussion that follows eq. 6). Using Theorem 2.3 (Ljung, 1999) for a quasi-stationary signal, we can prove 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_c is obtained using this method.

Wentzell *et al.* (1997) proposed a maximum likelihood estimation 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, it iteratively transforms

the model identified by PCA on unscaled data, until the maximum likelihood estimates of $x(t)$ are obtained. In contrast, the procedure we have 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_c 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.

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 we can estimate the error covariance matrix from the data along with the model. We describe an iterative algorithm for achieving this by combining PCA with a maximum likelihood estimation (MLE) method for obtaining an estimate of the error covariance matrix. We will assume that an initial estimate of the model constraint matrix, $\hat{A}^0$, is available. (Such an estimate can be obtained by applying PCA to the measured data). Using this initial model estimate, we compute the constraint residuals at each time instant as

$$r(t) = A^0 y(t) \tag{10}$$

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

$$\min_{\Sigma_\epsilon} N \log |\hat{A}^0 \Sigma_\epsilon (\hat{A}^0)^T| + \sum_{i=1}^{N} (r_i^T(t) (\hat{A}^0 \Sigma_\epsilon (\hat{A}^0)^T)^{-1} r_i(t)) \tag{11}$$

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 $r(1) \dots r(N)$) 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.

$$\hat{A}^0 \hat{\Sigma}_\epsilon (\hat{A}^0)^T = S_r \tag{12}$$

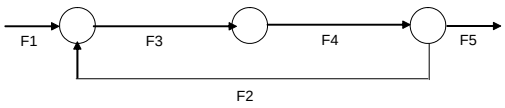

Fig. 1. Schematic of a flow process

Depending on the number of constraints and number of variables, it may or may not be possible to satisfy the above equation. Typically, the number of spatial relations m is usually less than n. In such cases, if we attempt to estimate all diagonal and off-diagonal elements of Σ_ϵ, multiple solutions that satisfy the above equation are obtained. One possibility is to assume that Σ_ϵ is diagonal, and estimate only the n diagonal elements corresponding to the measurement error variances. Even in this case, a non-degenerate estimate for Σ_ϵ is obtained only if $m(m+1) \geq 2n$. Other techniques have been proposed for estimating the measurement error covariance matrix, given the constraint model and the covariance matrix of constraint residuals (Romagnoli and Sanchez, 1999). However, these methods are not maximum likelihood estimates.

Assuming that the number of diagonal and off-diagonal elements of Σ_ϵ that we are estimating is less than or equal to $m(m+1)/2$, we can minimize (11). We can also impose lower bounds on the elements of Σ_ϵ that we are estimating, and solve the constrained optimization problem. Let us denote the estimate of measurement error covariance matrix obtained using the above method as $\hat{\Sigma}_\epsilon^0$. Note that this estimate has been obtained assuming that the model has been estimated exactly. We can use $\hat{\Sigma}_\epsilon^0$ to transform the data as described in the preceding section, and apply PCA on the transformed measurements to get an updated estimate of the constraint matrix. We can repeat the entire procedure 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.

6. SIMULATION RESULTS AND DISCUSSION

A flow process example shown in Fig. 1, has been chosen to test the proposed procedure. The above example has been chosen so that it satisfies the condition $m(m+1) > 2n$ (in the above example $m = 3$ and $n = 5$). We will assume that the measurement error covariance matrix is diagonal.

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

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

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_c, 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

$$\alpha_i = ||A_{.i}^T - A_{.i}^T \hat{A}^T (\hat{A}\hat{A}^T)^{-1} \hat{A}|| \qquad (13)$$

A consolidated measure of model estimation accuracy is given by

$$\alpha = \sum_i \alpha_i \qquad (14)$$

The above measure treats all bases sets for the row space of $\hat{A}$ as equivalent. 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.

The results obtained for the above example using PCA for different choices of data scaling and the proposed iterative method (denoted as IPCA) are presented in Table 2. 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, we can conclude that a good estimate of the model constraints is obtained using both PCA

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

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 obtained through PCA by more than 80%. The number of major iterations required for IPCA to converge was around ten, 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 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 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, we can alter the singular values, and it may make it difficult to determine the number of PCs to be retained/rejected. In other words, we may not be able to determine the number of constraints precisely by examining the singular values of the scaled data, unless we use the standard deviations of measurement errors for scaling. It may also be noted that if the data is auto-scaled, a worse model may be obtained compared to the case when the data is 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

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

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 our proposed method can be used even if errors in different variables are correlated, we simulated data 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 3 constraints, we can estimate at most 6 elements of the error covariance matrix. 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 non-zero 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.

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

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 non-zero 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.

We had stated that it may also be 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 std 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.

7. CONCLUDING REMARKS

In this paper, we have proposed an algorithm 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 eigenvalues obtained using PCA on the transformed data is proposed, which has a rigorous theoretical basis. As part of future research, the technique described here can be extended to methods which use PCA as an integral component such as PCR, PLS, and subspace based model identification.

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.

REFERENCES

Ljung, L. (1999). *System Identification: Theory for the User*. 2nd ed.. Prentice-Hall. New Jersey.

Morrison, D.F. (1967). *Multivariate Statistical Methods*. 2nd ed.. McGraw-Hill.

Romagnoli, J.A. and M.C. Sanchez (1999). *Data Processing and Reconciliation For Chemical Process Operation*. Academic Press.

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

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

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