University of Alberta

STATISTICAL MODELING AND SENSOR FAULT DIAGNOSIS USING PCA

By

Edward Y. Bai 

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in

Chemical Engineering

Department of Chemical and Materials Engineering

Edmonton, Alberta
Fall, 2005

Your file Votre référence
ISBN: 978-0-494-22425-0
Our file Notre référence
ISBN: 978-0-494-22425-0

University of Alberta

Library Release Form

Name of Author: **Edward Y. Bai**

Title of Thesis: **Statistical Modeling and Sensor Fault Diagnosis Using PCA**

Degree: **Master of Science**

Year This Degree Granted: **2005**

Edward Yaohua Bai
CME 268
Chemical and Materials
Engineering
University of Alberta
Edmonton, AB
Canada, T6G 2G6

Date:

Abstract

Principal Components Analysis (PCA) is used to reduce the dimensionality of the data matrix and capture the underlying variation and latent relationship among the variables. In recent years, PCA has become a popular tool for identifying linear models of processes from historical data. This technique has been used for process monitoring and fault diagnosis. Fault detection is usually carried out using SPE/T^2 statistics while contribution plots are used to identify the variables that are the primary causes of the fault. The success of fault-diagnosis methods depends upon:

(1) the accuracy of the process models we can get from the identification;

(2) the sensitivity of fault-detection techniques;

(3) the resolution quality of fault-isolation strategies.

(4) Robustness to model mis-match or non-linearity of the system

It is well know that the model obtained using PCA is optimal under the assumption that the measurement errors are identically, independently and normally distributed (iid normal), and the error covariance matrix $\Sigma_e = \sigma^2 \mathbf{I}$. However, this is seldom true in practice, as different sensors cannot be expected to have the same level of measurement noise. To circumvent this assumption but without guarantee, classical PCA typically applies auto-scaling to the process data to get unit-variance. A significant disadvantage of PCA, and also fault diagnosis using contribution plots, is that they both depend on the choice of data scaling.

Given the error covariance plus the assumption of normality, the linear model identification may be easily performed by a MLE approach. Unfortunately, the error covariance is usually unknown.

A technique known as *iterative PCA* (IPCA) has recently been proposed by Narasimhan and Shah (2004) for simultaneous estimation of both the model and error covariance matrix.

First, this thesis demonstrates through simulation, the significant advantages of IPCA over PCA in providing accurate estimates of both the model and error covariance matrix. Second, this thesis shows that the estimated model and error covariance matrix can be combined with the well established techniques of Data Reconciliation (DR) and Gross Error Detection (GED) for more accurate state estimation and sensor fault diagnosis. In particular, it is shown through simulation that significanlt improvement in sensor fault diagnosis can be obtained by using the generalized likelihood ratio (GLR) test approach as compared to the use of SPE/T^2 tests and contribution plots which are conventionally used with PCA based techniques. Furthermore, both the IPCA method as well as the GLR approach possess the characteristic of being invariant to scaling of the data. The following two important perspectives are obtained through this thesis.

- PCA and IPCA should be regarded as tools for model identification and not be bundled together with contribution plots for fault diagnosis. Significant improvement in diagnostic resolution can be obtained by using these models with well established statistical techniques such as likelihood ratio tests.
- The IPCA method and DR/GED techniques are shown to be complementary approaches for steady state processes. While IPCA is concerned with identifying a model and error covariance matrix from data, DR and GED are concerned with state estimation and fault diagnosis assuming the availability of the process model and error covariance matrix.

Acknowledgements

This project has been a challenging one and a rewarding learning experience for me. I owe my special thanks to my supervisor, Dr. Sirish Shah, who has been giving me continuous inspiration and encouragement ever since I joined his research group. I would like to thank him for the time and enthusiasm he has put into the supervision, his careful guidance, and the dynamic interaction during the entire research project for my MSc. program.

I am grateful to Dr. Biao Huang for his lectures in the graduate course ChE662 on system identification. I am also indebted to Dr. Shijie Liu for his support and encouraging advice when I worked with him before I joined the CPC group.

I owe much gratitude and thanks to Dr. Shankar Narasimhan for his help and guidance in this thesis. I enjoyed frequent inspiring discussions with him during his stay at the University of Alberta as a visiting professor. His ideas and suggestions indeed enriched this project.

I would also like to thank Dr. Scott Meadows for his excellent guidance in ChE646 - Process Dynamics and Computer Process Control. I am thankful to Dr. U. Sundararaj, Dr. Suzanne M. Kresta and Dr. Long Wu for the pleasant times I had with them as their teaching assistant. I would also like to express my thanks to Dr. Forbes and Dr. Yip San for the enthusiastic guidance they gave to me in the optimization course-ChE654, and to Prof. Hooper for his excellent lectures in STAT575.

During my MSc. Program, many of my colleagues in the CPC group gave me much generous help and friendship. They are Dr. Arun K. Tangirala, Dr. Weihua Li, Dr. Hancong Liu, Dr. Fardin Akbaryan, Dr. Liqian Zhang, Dr. Lisheng Hu, Fangwei Xu, Harigopal Raghavan, Zhengang Han, and Syed A. Imtiaz.

I would like to give my acknowledgement to the Department of Chemical and Materials Engineering, University of Alberta, for giving me the opportunity to fulfill my MSc. program and providing me with a wonderful academic environment and multiple accesses to research

resources. The friendly help and support that I have received from the teaching faculty and administrative staff in the department made me feel at home. I want to thanks Leanne and Theresa, their help and support was always there. I want to gratefully acknowledge Matrikon Inc. for providing me an excellent office environment and abundant resources, including the access to real industrial data and problems. At Matrkon, I received help from Yannis Faitakis, Rohit Patwardhan, Chris Shelton, Jianping Gao, Haitao Zhang, and Dave Shook.

I would like to gratefully acknowledge the financial support from the NSERC-Matrikon-ASRA Project. Last but not the least I want to express my thanks to my parents, my wife and daughter for their consistent support, care and love, and encouragement during the course of my graduate studies.

Table of Contents

Chapter 1

Introduction

1.1 Overview of Multivariate Statistical Process Control (MSPC)

Traditionally in the manufacturing industry, statistical quality control (SQC) is used for monitoring and controlling product quality. Because these are crucial factors contributing to profitability, laboratory tests are carefully scheduled to obtain current product quality data at an acceptable frequency. However, **by monitoring only the quality variables, we ignore the hundreds of process variables that are measured much more frequently than the product quality data** (Kourti, 2002). Process operation data may also be very useful for prediction and control of end product qualities. Further more, in the chemical industry, people are also concerned with the monitoring and enhancement of safety and process reliability, and the improvement of profitability, and the reduction of manpower costs. To accomplish these things, we need informative data from the process in addition to only quality variable data on the end products. Advances made in the areas of on-line instrumentation and data acquisition have made it possible to collect large amounts of data in the chemical process industry. Given that the data has a certain level of redundancy, it becomes possible to detect any abnormality and locate its source in the process.

To enhance process operation, we want to not only monitor the process in an efficient manner, but also successfully identify the source of abnormality that may result in any degradation of product quality, operation reliability and profitability, in order that we can respond accordingly by making any necessary correction to the process. Statistical process control (SPC) (Montgomery, 1996), multivariate statistical process control (MSPC), and Six Sigma (Hoerl, 1998) are some of the tools that have been applied to achieve these objectives. Univariate SPC charts are used to monitor key process variables to reveal any abnormalities. These statistical control charts include *Shewhart* (Shewhart, 1931), *cumulative sum* (CUSUM) (Page, 1954; Woodward and Goldsmith, 1964) and *exponentially weighted moving average* (EWMA) (Roberts, 1959; Hunter, 1986; Lucas and Saccucci, 1990) charts. Although these univariate control charts have been used in most industries, they are only appropriate under the assumption that each observed variable is independent of others. When looking at multivariate data, these methods will ignore the interaction between the correlated variables and therefore result in a misleading analysis.

In this context, MSPC and the associated statistical techniques are finding increased use in continuous and batch processes in chemical engineering. Through the extraction of information , MSPC enables us to gain knowledge of and insight into processes. Furthermore, MSPC can provide early warning of abnormal changes in process operation, helping us to identify the onset of potential plant faults (exchanger fouling, catalyst poisoning, etc.), equipment trips, actuator malfunctions, sensor bias, and unmeasured disturbances. In other words, multivariate statistical process monitoring and control **detects the existence, magnitude, and time of changes that cause a process to deviate from its desired operation** (Çinar and Ündey, 2002). Ordinary Least Square (OLS) Regression, Principal Component Analysis (PCA), Partial Least Square (PLS) and Canonical Variate Analysis (CVA) have been extensively applied in the field of chemometrics. Recently, they have been increasingly used, as key MSPC tools, in the chemical engineering area and have also been widely discussed by researchers.

A brief introduction is given below on common statistical techniques that are applied in MSPC. In the following chapters, the discussion will focus on PCA.

Ordinary Least Square (OLS) Regression

When we try to regress data block $\mathbf{Y}$ (a group of quality variables) to $\mathbf{X}$ (selected group of process variables whose values are known precisely) as

$$\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\varepsilon}_Y \qquad \ldots\ldots\ldots(1.1)$$

the unbiased maximum likelihood estimate of θ, if the matrix $\mathbf{X}^T\mathbf{X}$ is non-singular and the noise $\boldsymbol{\varepsilon}_Y \sim N(0, \Sigma)$, is given as

$$\hat{\theta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \qquad \ldots\ldots\ldots(1.2)$$

Partial Least Square (PLS)

PLS was first proposed by Wold (1966) in the field of econometrics. Later Geladi and Kowalski (1986a, 1986b) provided a detailed PLS algorithm. Phatak and Jong (1997) illustrated the geometry of two algorithms for carrying out PLS in both object and variable space. This technique is used in chemometrics and chemical engineering for soft sensor development (Jansson *et al.,* 2002), process monitoring, and fault diagnosis (Wangen and Kowalski, 1988; MacGregor *et al.*, 1994a, Kourti *et al.*, 1995; Lakshminarayanan, 1997; Westerhuis and Coenegracht, 1997; Zhang, 2000).

PLS also stands for Projection on Latent Structures. When the matrix $\mathbf{X}^T\mathbf{X}$ is singular or ill-conditioned, PLS finds an optimum pair of latent variables both in $\mathbf{X}$ and $\mathbf{Y}$ such that these

transformed variables have the largest covariance. The first pair of latent variable vectors $\{t_1 = \mathbf{X}p_1;\ u_1 = \mathbf{Y}q_1\}$ is calculated so that the following covariance

$$\max_{p_1, q_1} (\mathbf{X}p_1)^T (\mathbf{Y}q_1) \qquad \ldots\ldots\ldots (1.3)$$

can be maximized with constraints $|p_1| = 1$ and $|q_1| = 1$. It turns out that p_1 is the first eigenvector of matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ and q_1 is the first eigenvector of matrix $\mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{Y}$.

Canonical Variate Analysis (CVA)

Similar to PLS, CVA normalizes the latent variable vectors u_i and t_j in **X** and **Y** data spaces. So, CVA maximizes the correlation between the pairs of these two latent variable vectors. CVA easily results in significant dimensional reduction of the data. This is because it focuses only on latent variables in the sequential order of the significance of their correlations. The effect of the excitation magnitude in data **X** or **Y** is eliminated by normalization. The use of CVA in regression and system identification can be found in the literature (Larimore *et al.*, 1984 & 1990; Schaper *et al.*, 1994; Lakshminarayanan, 1997; Burnham *et al.*, 1996; Dehon and Filzmoser, 2000).

Principal Component Analysis (PCA)

PCA is probably the most commonly used technique for dimensionality reduction. The introduction of PCA can be traced back to Pearson (1901) and Hotelling (1933). PCA has become a popular modeling technique to extract information from process data by relating process variables. In this context, PCA *scores*, which are linear conbination of physical variables, can represent a process effectively in a reduced subspace. The method has been found useful in many applications, such as data compression, image analysis, visualization, pattern recognition, chemometrics. In chemical engineering, PCA is generally used for outlier detection, data filtering, and data smoothing or reconciliation (Kramer and Mah, 1994), regression and time series prediction (Filzmoser, 2001), gross error detection (Tong and Crowe, 1995), process monitoring (Kresta *et al.*, 1991) and fault diagnosis (MacGregor *et al.*, 1994a & 1994b; Dunia *et al.*, 1996).

PCA is different from PLS and CVA in that PCA does not differentiate between data sets **X** and **Y**. It is applied to one data set that contains all the process variables concerned in the problem. We will use notation **Y** to represent the whole data set for the sake of consistency with the following chapters, whereas notation **X** is often used in the literature.

PCA is performed on the normal operating data (training data) enabling us to obtain a pair of models: a *process model* and its complement, which is defined as a *constraint model.* Abnormal events are detected if the measurements deviate from the region of normal operation in the *principal component space* (PCS) or in the *residual space* (RS). The scores and residuals are

often plotted in univariate SPC charts. PCS forms the *process model* and RS forms the *constraint model.*

Numerous extensions of PCA have been devised to meet various requirements in practical use. Multiway PCA allows the analysis of data from batch processes (Nomikos and MacGregor, 1994; Çinar and Ündey, 2002). Hierarchical PCA permits easier modeling and interpretation of a large matrix by decomposing it into smaller matrices (Wold *et al.*, 1996; MacGregor, 1994b; Westerhuis *et al.*, 1998). Dynamic PCA identifies both spatial and temporal relationships in the data matrix augmented by time-lagged variables (Kresta *et al.*, 1991; Ku *et al.*, 1995; Tsung, 2000; Li and Qin, 2001). Nonlinear PCA reveals non-linear relationships between variables (Kramer, 1991; Schölkopf, 1998; Jia, *et al.*, 2000; Yu, 2002; Shi and Tsung, 2003). Multiscale PCA (MSPCA) indicates the capabilities of modeling and monitoring process at different frequency bands. MSPCA using wavelets is used for data de-noising and reducing autocorrelation in the data (Bakshi, 1998; Luo *et al.*, 1999; Misra, 2002). Robust PCA estimates the eigenvalues and eigenvectors that are tolerant with respect to possible outliers (Li, 1985; Rouseeuw, 1999; Skočaj, 2002). Recursive PCA updates the model continuously on-line (Li *et al.*, 2000); Similarly, on-line adaptive PCA updates the model using EWMA (Wold, 1994).

The Role of PCA in System Identification and Fault Detection and Isolation

In general, we have to obtain the *model* first and then perform fault diagnosis procedures accordingly. The estimation of the residual space, which is crucial in FDI, depends on the appropriate process modeling. There are many approaches to process modeling; and overviews on this topic are readily available in the literature. Figure 1.1 outlines the role that PCA plays in process modeling or system identification.

Venkatasubramanian (2003) has provided a thorough overview of process fault detection and diagnosis (FDD). Figure 1.2 shows how PCA fits into the hierarchical Classification of diagnostic algorithms.

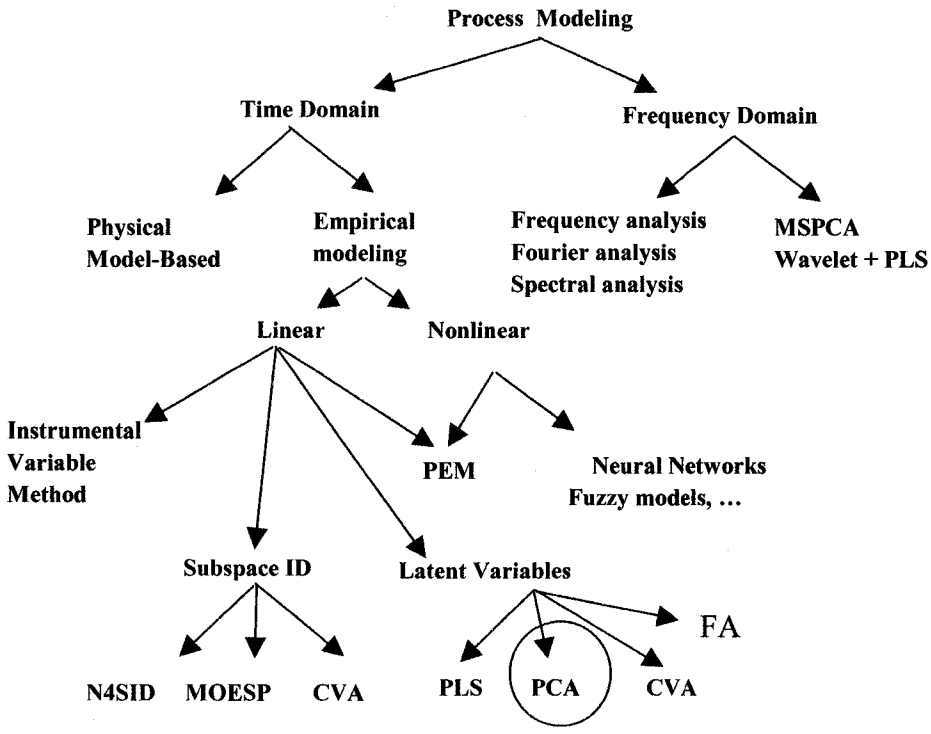


Figure 1.1 A summary of various modeling approaches

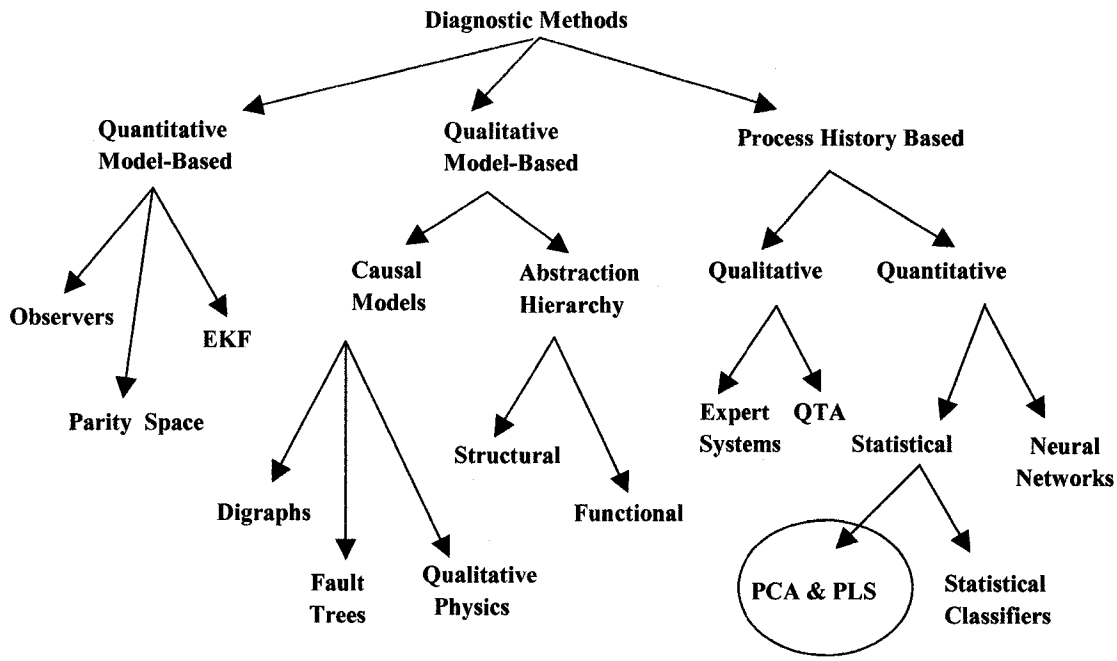


Figure 1.2 A summary of various FDI approaches

1.2 The Scope of this Thesis

Motivation

Although process model estimation using conventional PCA is generally fairly good, it depends on scaling and therefore neither necessarily give an optimal nor an unbiased estimate. Thus, there is a demand for a scaling invariant optimal approach for process modeling using PCA. This thesis studies the efficacy of a recently proposed IPCA technique (Narasimhan and Shah, 2004) for simultaneously estimating both the process model and error covariance matrix.

This thesis studies in detail through Monte Carlo simulations the advantages offered by the IPCA method over the PCA method for steady state linear model identification. Secondly, it is shown that the well known SPE test and corresponding contribution plots do not always give a

satisfactory resolution to the fault isolation problem. Therefore, we develop simple and more sensitive alternatives to SPE and its contribution plot based on likelihood ratio statistical test for sensor fault diagnosis. In this context, the model and error covariance matrix estimates provided by the IPCA method are shown to be extremely useful in improving diagnostic resolution of sensor faults. Through this, the link between PCA and IPCA methods and the well researched area of Data Reconciliation and Gross Error Detection (Sanchez and Romagnoli, 2000; Narasimhan and Jordache, 2000) is clearly established in this thesis.

Thesis organization

Following this introduction, Chapter 2 starts with the review of basic problem formulation and algorithm of the PCA method. Related topics, such as model order determination and scaling are included in the preparatory discussion to provide background for the forth coming discussion. For a better understanding of PCA as a model ID method, a detailed comparisons between PCA and TSL, PCA and regular LS regression are provided. Later in the same chapter, a recently proposed method, *iterative PCA* (or IPCA), is described which can identify more accurate process models. Chapter 3 provides a detailed simulation study comparing the advantages of the IPCA method over the PCA method for steady state model identification and for state estimation. The link between state estimation using IPCA and *Data Reconciliation* (DR) is established in chapter 4. Then, in chapter 5, methods for FDI using PCA model are introduced in detail and the limitations of SPE-based Q statistics are discussed quantitatively. Squared weighted residual (SWR) and generalized likelihood ratio (GLR) are proposed as improved alternatives for fault detection and isolation respectively. The superior performance of the new combined FDI strategy (IPCA-SWR-GLR) is also demonstrated via Monte Carlo simulations.

Further discussion of data properties and data pre-processing is provided in Chapter 6, followed by some suggestions on practical applications for process modeling, monitoring and fault diagnosis. Many related topics, such as variable selection, dealing with colored noise, fault detectability, SWR threshold training, are also included where necessary. The thesis ends by presenting a set of firmly established conclusions and by enumerating a list of topics of interest for future research.

Chapter 2

PCA, Optimally Scaled PCA and Iterative PCA for Model Identification

2.1 Introduction

When there is a high degree of colinearity, i.e., the process variables are strongly correlated, the rank of an observed data set is much less than the number of variables. *Principal Component Analysis* (PCA) is a multivariate statistical method in chemometrics that has been intensively studied and widely used for the rank reduction of the observed data and, at the same time, reveal underlying relationships among variables. The covariance structure in the data can be explained in a reduced dimensional space through an orthogonal set of latent variables, i.e., a set of linear combinations of the original variables. More precisely, PCA involves finding one direction such that the projection of the data in that direction explains the greatest variability of the data, followed by finding, in the same manner, the next direction that is orthogonal to all the previous ones and so on. The values of the latent variables are *scores* read from these projections. We also call these values *score vectors* (with the same length as the number of observations). The number of latent variables is the same as the total number of variables. Nevertheless, due to the dependency and colinearity, it is usually the case that much of the variation can be captured by only a small number of latent variables. This part of the latent variables constitutes a set of *principal components* (PC) or factors. Figure 2.1 shows how 3-dimensional co-linear data can be represented in a reduced 2-dimensional space using only 2 principal components.

All latent variables (including principal components as part of them) are connected to the original variables by linear combinations, with the coefficients represented as a group of *loading vectors*. These loading vectors are singular vectors of the data matrix. On one hand, the loading vectors corresponding to a group of significantly large singular values explain the principal components and span the *principal component subspace* (PCS) for the data (i.e., a k-dimensional hyperplane). On the other hand, the loading vectors corresponding to the remaining group of small singular values span the *residual subspace* (RS) or *null space*. We call this subspace the *process constraint model*. In this context, PCA is a method for model extraction or system identification. An m-dimensional linear constraint model means that the data conforms with m constraints (proof

is given in Appendix A). Principal components in PCS capture the variance information and are typically used for dimensionality reduction (data compression), while constraint models are typically used for data filtering or data reconciliation (DR) as well as for fault diagnosis. In this thesis, **the terminology "model" always means the constraint model**, unless defined otherwise. This chapter will discuss PCA technique from a modeling perspective.

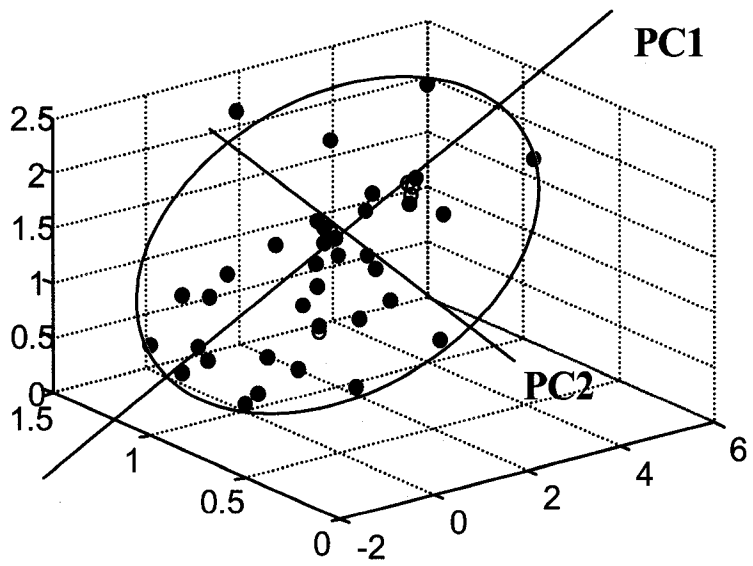


Figure 2.1 PCA expresses 3-dimensional data in a reduced 2-dimensional plane spanned by orthogonal vectors (or principal components) PC1 and PC2

2.2 Introduction to PCA

Principal Component Analysis deals with measurement error and in this respect it belongs to the wide group of measurement error models. In the linear measurement error type of model a wide number of eigenvalue-based methods have been developed. In this section we will start with the formulation of the general problem, where PCA falls in this classification and its connection with other eigenvalue-based methods.

2.2.1 General Form of Linear Measurement Error Model

The basic form of the linear measurement error model is as follows: Let there be N individual observation of n $(<N)$ variables. The measurements are grouped into an $n \times N$ matrix $\mathbf{Y}$ of rank n

and are subject to measurement error. Let **X** be the conformable matrix of true values and **E** be the matrix of measurement errors,

$$\mathbf{Y} = \mathbf{X} + \mathbf{E}$$

The elements of **X** can be either stochastic or non-stochastic. The rows of **E** are assumed i.i.d. with zero mean and covariance matrix Σ. The model is concerned with the rank of **X**, and it is assumed that there exists some linear relationship between the columns of **X**. This relationship can be expressed in two forms. Under the Principal Relation (PR) specification, there exists an $m \times n$ matrix **A**, with $m < n$, of full row rank such that

$$\mathbf{AX} = \mathbf{0}$$

The columns of **X** are restricted to lie in a $(n-m)$ dimensional subspace.

In a complementary Principal Factor (PF) specification, the restriction can be imposed by expressing **X** as a product of two factors. In this form it is assumed that there exists a matrix **T** $(k \times N)$, and a matrix **P** $(n \times k)$ with k $(< n)$ such that

$$\mathbf{X} = \mathbf{PT}$$

where the rows of **T** are called the factors of **X**.

Principal Relations and Principal Factor are equivalent in the sense that, they are able to impose the same restrictions on **X** when $n = k + m$.

2.2.2 Principal Component Analysis

Under the above classification PCA is formulated in PF form. Only a few assumptions are made in PCA analysis.

- True signal **X** is rank deficit and there exists this relation: $\mathbf{X} = \mathbf{PT}$
- The measurement error, **E** and the true signal **X** are independent of each other.
- There exists good excitation and measurement noise is small, i.e. sufficient signal to noise ratio.

Under these assumptions PCA tries to estimate **X** by minimizing **E** in least square sense. With further assumption that all the measurement errors are iid-normal in variable direction $(i.e.\ \Sigma = \sigma^2\mathbf{I})$ it can be shown that PCA gives Maximum Likelihood Estimate (MLE). The detailed derivation of PCA is described below.

Maximizing the Variability in the Data to Find PCs

Suppose that we have data (or measurement) matrix $\mathbf{Y}_{nxN}$, where N represents the number of samples and n represents the number of process variables. We will assume, for this formulation, that the data mean has already been removed and data variance is scaled to unity. In other words, the data is auto-scaled.

$$\mathbf{Y} = \begin{bmatrix} y_1(1) & y_1(2) & \cdots & y_1(N) \\ y_2(1) & y_2(2) & \cdots & y_2(N) \\ \vdots & \vdots & \vdots & \vdots \\ y_n(1) & y_n(2) & \cdots & y_n(N) \end{bmatrix} = \mathbf{X} + \mathbf{E} \qquad \ldots\ldots\ldots (2.1)$$

In PCA, the values (scores) for the first latent variable (the first PC) form a score vector:

$$t_1 = v_1^T \mathbf{Y}$$

The scores are linear combinations of the original data that should account for the maximum variance in the data. Here the elements of the first loading vector v_1 are coefficients of the linear combination. The score vector t_1 is a row vector with N elements.

Solving the following optimization problem for v_1:

$$\max_{v_1} t_1 t_1^T = v_1^T \mathbf{Y}\mathbf{Y}^T v_1 \qquad \ldots\ldots\ldots (2.2a)$$

$$s.t.\ v_1^T v_1 = 1$$

combining the last two equations via the Lagrangian gives:

$$L = v_1^T \mathbf{Y}\mathbf{Y}^T v_1 - \lambda\left(v_1^T v_1 - 1\right)$$

Setting the partial derivative with respect to v_1 to zero gives:

$$\frac{\partial L}{\partial v_1} = 2\mathbf{Y}\mathbf{Y}^T v_1 - 2\lambda v_1 = 0$$

$$\mathbf{Y}\mathbf{Y}^T v_1 = \lambda v_1$$

Thus, v_1 is the eigenvector of $\mathbf{Y}\mathbf{Y}^T$ with λ as the corresponding eigenvalue. If we look at the second derivative:

$$\frac{\partial^2 L}{\partial v_1^2} = 2\mathbf{Y}\mathbf{Y}^T - 2\lambda\mathbf{I}$$

For maximization problem, we require the condition:

$\partial^2 L/\partial v_1^2 \le 0 \Rightarrow e^T\left(\mathbf{YY}^T - \lambda\mathbf{I}\right)e \le 0$, where $e^T e = 1$. i.e., $e^T\left(\mathbf{YY}^T\right)e \le \lambda$.

This is true only when λ is the largest eigenvalue of the matrix $\mathbf{YY}^T$. We know that λ /(N-1) is nothing but the variance of score t_1, which explains the variability of data $\mathbf{Y}$ in the direction defined by v_1. Therefore, the corresponding eigenvector v_1 is the solution of this problem, i.e., v_1 is the first PC.

The information unexplained by the first PC is:

$$\Delta_1 = \mathbf{Y} - v_1 t_1$$

Δ_1 is called the residual or the deflated data matrix. To find the second latent variable v_2, we solve the optimization problem similarly:

$$\max_{v_2} t_2 t_2^{\mathrm{T}} = v_2^T \Delta_1 \Delta_1^T v_2 \qquad \ldots\ldots\ldots (2.2b)$$

$$s.t.\ \ v_2^T v_2 = 1$$
$$\&\quad v_2^T v_1 = 0$$

It turns out that the eigenvector associated with the second largest eigenvalue is the solution for the second latent variable v_2, which is also called the second PC if the eigenvalue is considered large enough. This procedure is continued until n latent variables are obtained. Given certain correlation (redundancy) in data $\mathbf{Y}$, the first 'k' latent variables (or principal components) (where $k \prec n$) are able to capture most of the variability in $\mathbf{Y}$. This part of the variability commonly arises from the true underlying signals. The remaining m ($m = n - k$) latent variables represent the residuals of the process constraint equations. They capture the variability that arises from noise. All n latent variables collectively explain the same amount of variability as in data $\mathbf{Y}$.

If we use $\mathbf{YY}^T/(N-1)$ instead of $\mathbf{YY}^T$, its orthogonal eigenvectors are equal to the loading matrix $\mathbf{V} = (v_1, v_2, \ldots, v_n)$, and eigenvalue λ_i is the variance of score t_i. Here, for convenience of discussion, we partition the loading matrix into two parts, $\mathbf{P}$ and $\mathbf{B}$:

$$\mathbf{V} = [\mathbf{P}\ \mathbf{B}]$$

Here $\mathbf{P} = (v_1, v_2, \ldots, v_k)$ represents the first k principal loading vectors (PCs) and $\mathbf{B} = (v_{k+1}, v_{k+2}, \ldots, v_n)$ represents the remaining n - k loading vectors. The partition is shown below:

$$\frac{\mathbf{Y}\mathbf{Y}^T}{N-1} = \mathbf{V}\Lambda\mathbf{V}^T = [\mathbf{P}\ \mathbf{B}]\begin{bmatrix}\Lambda_p & 0\\ 0 & \Lambda_b\end{bmatrix}\begin{bmatrix}\mathbf{P}^T\\ \mathbf{B}^T\end{bmatrix} \qquad \ldots\ldots\ldots (2.3)$$

We have:

$$\mathbf{Y} = \mathbf{P}\mathbf{T}_p + \ \mathbf{B}\mathbf{T}_b = \hat{\mathbf{X}} + \hat{\mathbf{E}} \qquad \ldots\ldots\ldots (2.4)$$

where $\mathbf{T}_p = \mathbf{P}^T\mathbf{Y}$

$$\mathbf{T}_b = \mathbf{B}^T\mathbf{Y}$$

$$\hat{\mathbf{X}} = \mathbf{P}\mathbf{T}_p = \sum_{i=1}^{k} v_i t_i$$

$$\hat{\mathbf{E}} = \mathbf{B}\mathbf{T}_b = \sum_{i=k+1}^{n} v_i t_i$$

$$\Lambda_P = \mathrm{cov}(\mathbf{T}_p),\ \Lambda_b = \mathrm{cov}(\mathbf{T}_b) \qquad \ldots\ldots\ldots (2.5)$$

$\mathbf{T}_p$ is the principal component score matrix ($k \times N$), which describes the values of variables in the transformed $n \times k$ basis space spanned by **P**. Here k is chosen such that there is no significant process information left in $\hat{\mathbf{E}}$, and $\hat{\mathbf{E}}$ is expected to contain only the random error. Thus, the group of orthogonal principal loading vectors (or PCs), **P**, forms the *principal component subspace* (PCS), which describes the systematic variation in the data. Adding extra PCs to the PCs ends up fitting the random error.

Estimation of the Constraint Model From PCA

If we regard the columns of **P** (the first k principal loading vectors) as a basis for the true data vectors in **X**, then the constraint model can be obtained from the remaining n-k loading vectors as $\mathbf{A} = \mathbf{B}^T$. This follows from the orthonormality of the loading vectors. Here **B** is obtained by choosing the last m latent variables in **V** such that **B** represents only the variability of random errors. **B** forms the *constraint model* denoted as $\mathbf{A} = \mathbf{B}^T$. For error free data **X**, we have $\mathbf{AX} = \mathbf{0}$, because: From equation 2.4,

$$\mathbf{Y} = \hat{\mathbf{X}} + \hat{\mathbf{E}} = \mathbf{P}\mathbf{P}^T\mathbf{Y} + \mathbf{B}\mathbf{B}^T\mathbf{Y}$$

Multiplying from the left by $\mathbf{A}=\mathbf{B}^T$:

$$\mathbf{AY} = \mathbf{B}^T\mathbf{PP}^T\mathbf{Y} + \mathbf{B}^T\mathbf{BB}^T\mathbf{Y} = \mathbf{B}^T\hat{\mathbf{E}} \qquad \ldots\ldots\ldots (2.5a)$$

Notice that:

$$\mathbf{A}\hat{\mathbf{X}} = \mathbf{B}^T\mathbf{PP}^T\mathbf{Y} = \mathbf{0} \qquad \ldots\ldots\ldots (2.5b)$$

This reveals that $\mathbf{A}$ is the estimate of the regression model, and $\mathbf{A}\hat{\mathbf{E}} = \mathbf{B}^T\hat{\mathbf{E}} = \hat{\mathbf{r}}$ is the estimate of the true constraint residuals. A precise proof is given in Appendix A.

Use of Singular Value Decomposition for PCA

In implementing the PCA algorithm, the Singular Value Decomposition (SVD) is applied to the observed data set $\mathbf{Y}^T$:

$$\frac{1}{\sqrt{N-1}}\mathbf{Y}^T = \mathbf{USV}^T$$

$\mathbf{U} \in R^{NxN}$ and $\mathbf{V} \in R^{nxn}$ are unitary matrices, and $\mathbf{S} \in R^{Nxn}$ is diagonal matrix containing nonnegative real singular values in decreasing magnitude. The singular values are equal to the square roots of the eigenvalues of the covariance matrix of $\mathbf{YY}^T/(N-1)$. The loading vectors are the right-singular vectors - the orthogonal columns of matrix $\mathbf{V}$.

PCA Becomes MLE Under Certain Conditions

The series of optimization problems for PCA shown in equation 2.2a and equation 2.2b can also be **equivalently formulated** as the following optimal estimation problem shown in equation 2.6, where the sum of estimation errors from all the variables is minimized (Hastie and Stuetzle, 1989):

$$\left\{\hat{\mathbf{P}}, \hat{t}_j\right\}_{PCA} = \underset{\hat{\mathbf{P}}, \hat{t}_j}{\operatorname{argmin}} \sum_{j=1}^{N} \left(y_j - \hat{x}_j\right)^T \mathbf{I} \left(y_j - \hat{x}_j\right) \qquad \ldots\ldots\ldots (2.6)$$

$$s.t. \quad \hat{x}_j = \hat{\mathbf{P}}\hat{t}_j,\ \hat{t}_j = \hat{\mathbf{P}}^T y_j \text{ and } \hat{\mathbf{P}}^T\hat{\mathbf{P}} = \mathbf{I}$$

where y_j and $\hat{x}_j$, which are $n\times1$ vectors, are the jth measured and estimated observations respectively, and $\hat{t}_j$ is a $k\times1$ vector of the estimated principal component score at observation y_j. $\hat{\mathbf{P}}$ is the $n\times k$ loading matrix formed by selected k principal components (PCs). An identity-normalizing matrix is used in equation 2.6. From this equation we can easily find that the PCA approach becomes a maximum likelihood estimate (MLE), under the assumption:

$$\text{cov}(\mathbf{E}) = \Sigma_e = \sigma^2\mathbf{I}$$

i.e., PCA implicitly assumes an equal noise contribution in all variables to be MLE.

2.2.3 Similar Eigenvalue-based Method: Total Least Squares (TSL)

The eigenvalue-based method that deserves discussion is *Total Least Squares* (TLS) also known as *Orthogonal Regression* (OR). We will discuss it here because of its close resemblance to PCA. One way to estimate regression coefficients is to fit a hyperplane such that, the sum of the squared distances from the observations to that hyperplane are minimal (figure 2.1a).

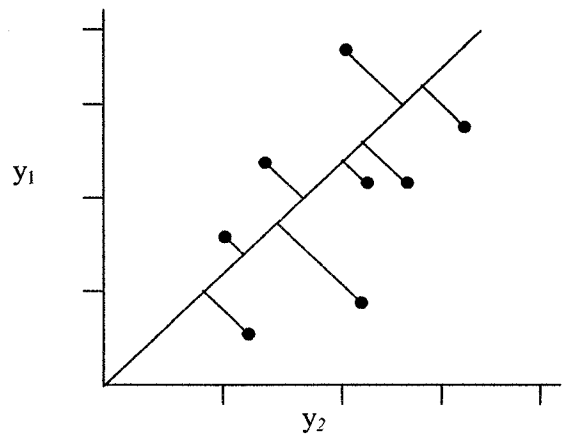


Figure 2.1a Orthogonal Regression

This contrasts with the usual way regression is performed, where distances are measured vertically in the direction of the dependent variable. It will appear below that OR leads to an eigenvalue solution (Pearson, 1901). *Orthogonal Regression* is formulated in the principal relations (PR) form. Let us start from a regression point of view where $\mathbf{X}_I$ is a matrix of independent variables and $\mathbf{X}_D$ is a matrix of dependent variables. Both $\mathbf{X}_D$ and $\mathbf{X}_I$ are corrupted with measurement noise. For a typical observation j this is given as,

$$y_{I,j} = x_{I,j} + \varepsilon_{I,j}$$
$$y_{D,j} = x_{D,j} + \varepsilon_{D,j}$$

or in a compact form, $y_j = x_j + \varepsilon_j$

In a regression setting we assume there exists a relation between $x_{D,j}$ and $x_{I,j}$.

$$x_{D,j} = \beta^T x_{I,j}$$
$$\Rightarrow \begin{bmatrix} 1 & -\beta^T \end{bmatrix} \begin{bmatrix} x_{D,j} \\ x_{I,j} \end{bmatrix} = 0 \qquad \ldots\ldots\ldots (2.7)$$
$$\Rightarrow a^T x_j = 0$$

So the objective here is to minimize the distance from the N observations $y_j = \begin{bmatrix} y_{D,j} & y_{I,j} \end{bmatrix}^T$ to the hyperplane where the rank deficit matrix $x_j, (j = 1.....N)$ lies on. The constraint, equation 2.7 is also included in the objective function and for a single observation we have to minimize

$$\frac{1}{2}\left(y_j - x_j\right)^T \left(y_j - x_j\right) + \lambda_j \left(a^T x_j\right) \quad \ldots\ldots\ldots (2.8)$$

where λ is the Lagrange multiplier.

Hence the first order condition, upon differentiating with respect to x_j is:

$$y_j - x_j = \lambda_j a \quad \ldots\ldots\ldots (2.9)$$

Although the first element of a has been normalized to 1, without loss of generality we can scale it to $a^T a = 1$. Then pre-multiplication by a^T gives:

$$\lambda_j = a^T \left(y_j - x_j\right) = a^T y_j \quad \ldots\ldots\ldots (2.10)$$

Combining equation 2.9 and equation 2.10 gives:

$$x_j = \left(I - aa^T\right) y_j$$

The squared distance from y_j to the hyperplane x_j can be expressed as bellow, following Equation 2.10,

$$\left(y_j - x_j\right)^T \left(y_j - x_j\right) = \lambda_j^2 = a^T y_j y_j^T a$$

and the sum squared distance for N observations is $(N-1)a^T \mathbf{S} a$ where $\mathbf{S}$ is the sample covariance matrix. Next, a is found by minimizing this subject to $a^T a = 1$. The Lagrange function corresponding to this problem leads to the eigenvalue equation:

$$(\mathbf{S} - \mu I)a = 0$$

so a is the eigenvector of **S.** In the optimum, $a^T \mathbf{S} a = \mu a^T a = \mu$. Therefore, the eigenvector corresponding to the smallest eigenvalue should be taken (Wansbeek, T and Meijer, E, 2000).

2.2.4 Connection Between TSL and PCA

In PCA we solved the following eigenvalue problem for **largest** eigenvalues.

$$\left(\mathbf{Y}\mathbf{Y}^T - \lambda \mathbf{I}_n\right)\mathbf{P} = \mathbf{0} \quad \ldots\ldots\ldots (2.11)$$

We can compare this with the Total Least Squares (or Orthogonal Regression) by putting $\Sigma = \sigma^2 \mathbf{I}_n$, $m = 1$ (rank of **P**) and therefore **P**=a. Now all variables are dealt with in a symmetrical way and we have to solve

$$\left(\mathbf{Y}\mathbf{Y}^T - \lambda \mathbf{I}_n\right)a = \mathbf{0} \quad \ldots\ldots\ldots (2.12)$$

for **minimal** λ . Distance is measured perpendicular to the $(n-1)$ dimensional hyperplane $a^T\mathbf{X}=0$. The complement of a is formed by the set of eigenvectors corresponding to the $(n-1)$ largest roots of $\mathbf{YY}^T$, which are exactly the principal component solution discussed earlier (Wansbeek, T and Meijer, E, 2000).

Total Least Squares can also be implemented by:

Minimizing Residuals in the Residual Space (RS)
Find orthogonal directions **V** so that the projections of **Y** represent minimum variability in the data:

$$\min_{v_i} J_{PCA} = \left(v_i^T\mathbf{Y}\right)\left(v_i^T\mathbf{Y}\right)^T \qquad \ldots\ldots\ldots (2.13)$$
$$s.t. \quad v_i^T v_i = 1$$

The solution turns out to be as same as equation 2.12, and also as same as **V** = **[P B]** in equation 2.3.

2.2.5 Least Squares Regression (LSR)

Multivariate linear regression deals with the problem of identifying the linear relationship that relates a given set of dependent variables and independent variables. Let the *n* variables be partitioned as *m* dependent variables denoted as $\mathbf{X}_D$ and the remaining *n-m* independent variables be denoted as $\mathbf{X}_I$. Let us assume that linear relations between the dependent and independent variables exist which are defined by

$$\mathbf{X}_D = \boldsymbol{\beta}\mathbf{X}_I \qquad \ldots\ldots\ldots (2.14a)$$

Given a sample of *N* measurements of the dependent variables arranged as a $m \times N$ data matrix $\mathbf{Y}_D$, which are related to the true values of the dependent variables as

$$\mathbf{Y}_D = \mathbf{X}_D + \boldsymbol{\varepsilon}_D \qquad \ldots\ldots\ldots (2.14b)$$

and a corresponding sample of *N* measurements of the independent variables $\mathbf{Y}_I$, which are equal to the true values $\mathbf{X}_I$ (that is, they do not contain any error), the objective is to estimate the $m \times (n-m)$ regression matrix $\boldsymbol{\beta}$.

The regression matrix is obtained by minimizing the following objective function

$$\min_{\boldsymbol{\beta}} \left(\mathbf{Y}_D - \boldsymbol{\beta}\mathbf{X}_I\right)^T\left(\mathbf{Y}_D - \boldsymbol{\beta}\mathbf{X}_I\right) \qquad \ldots\ldots\ldots (2.14c)$$

$$\hat{\boldsymbol{\beta}} = \mathbf{Y}_D\mathbf{X}_I^T\left(\mathbf{X}_I\mathbf{X}_I^T\right)^{-1} \qquad \ldots\ldots\ldots (2.15)$$

It should be noted that the solution given by equation (2.15) is not in standard form since our data matrix is arranged such that the columns correspond to the different sample measurement vectors. The following assumptions are made in obtaining the above LS solution.

- The formulation of the problem needs a priori knowledge of which variables should be assigned in $\mathbf{Y}_D$ and which variables in $\mathbf{X}_I$.
- The measurements of $\mathbf{X}_I$ do not contain any errors.
- $\mathbf{X}_I\mathbf{X}_I^T$ has full rank (this assumption is needed for the implementation of the algorithm).

If we make the following additional assumption that the errors in measurements of different samples of $\mathbf{y}_D$ are independently and identically normally distributed with mean zero and known covariance matrix $\Sigma_{\varepsilon_D} = \sigma^2\mathbf{I}$, then the following MLE based formulation (2.16) reduces to the LS objective function 2.14c. In fact, if Σ_{ε_D} has full rank, the solution 2.15 can be proved to be a MLE.

$$\min_{\boldsymbol{\beta}} \left(\mathbf{Y}_D - \boldsymbol{\beta}\mathbf{X}_I\right)^T \boldsymbol{\Sigma}_{\varepsilon_D}^{-1}\left(\mathbf{Y}_D - \boldsymbol{\beta}\mathbf{X}_I\right) + N\log\left|\Sigma_{\varepsilon_D}\right| \qquad \ldots (2.16)$$

2.2.6 Comparison Between LSR and PCA

In LS Regression the assumption that the measurements of $\mathbf{X}_I$ do not contain any errors rarely holds. Consider the case when the measurements of the independent variables $\mathbf{Y}_I$ contain errors and are related to their corresponding true values by

$$\mathbf{Y}_I = \mathbf{X}_I + \boldsymbol{\varepsilon}_I \qquad \ldots\ldots (2.17)$$

We can still choose to use solution 2.15, however the $\hat{\boldsymbol{\beta}}$ is no longer a MLE as supposed to be. Another option illustrated below is to estimate the constraint model $\mathbf{A}$ using PCA and then transform it into a regression form.

From equation 2 .14a, 2.14b, and 2.17 we have

$$\begin{bmatrix}\mathbf{I} & -\boldsymbol{\beta}\end{bmatrix}\begin{bmatrix}\mathbf{X}_D \\ \mathbf{X}_I\end{bmatrix} = \begin{bmatrix}\mathbf{I} & -\boldsymbol{\beta}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_D - \boldsymbol{\varepsilon}_D \\ \mathbf{Y}_I - \boldsymbol{\varepsilon}_I\end{bmatrix} = \mathbf{0} \qquad \ldots\ldots (2.17a)$$

$$\Rightarrow \begin{bmatrix}\mathbf{I} & -\boldsymbol{\beta}\end{bmatrix}\mathbf{Y} = \begin{bmatrix}\mathbf{I} & -\boldsymbol{\beta}\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_D \\ \boldsymbol{\varepsilon}_I\end{bmatrix} \to \mathbf{0} \text{ or} \qquad \ldots\ldots (2.17b)$$

$$\Rightarrow \begin{bmatrix}\mathbf{I} & -\boldsymbol{\beta}\end{bmatrix}\mathbf{X} = \mathbf{0} \qquad \ldots\ldots (2.17c)$$

Comparing these with equation 2.5a and 2.5b, we can transform **A** to a regression form by

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\beta}_{PCA} \end{bmatrix} \quad \ldots\ldots (2.18)$$

The matrix $\boldsymbol{\beta}_{PCA}$, can be obtained, if required, from **A** by performing linear operations such that the columns corresponding to the specified independent variables becomes an identity matrix of rank *n-m*. The matrix $\boldsymbol{\beta}_{PCA}$ then corresponds to the last *m* columns of this transformed matrix. It should also be noted that if the independent variables are specified incorrectly (i.e. if dependencies exist between these variables) then the columns of **A** corresponding to the "independent" variables will be singular or ill-conditioned. Table 2.1 gives a summary of the properties of LSR and PCA approaches for linear regression.

Method	Advantage	Disadvantage	Similarity
LSR	• Just requires that Σ_{ε_D} has full rank to be a MLE	• Pre-defined model structure • Require **X** to be full rank • Assumes no errors in **X**, which is not likely to be true	• Both assume no temporal correlation in ε_D and ε_I • Both may give a MLE estimate for β under certain conditions
PCA (TLS)	• No requirement for full-rank **X** • Automatically reveals the relationships between all variables • Allows errors in all variables • Allows Σ_{ε_D} to be singular	• To obtain MLE, one needs the assumption that $\text{cov}(\mathbf{E}) = \Sigma_e = \sigma^2\mathbf{I}$	

Table 2.1 Comparing LSR and PCA from a linear regression point of view

2.3 Determination of Model Order - Optimal Dimension of PC's

As mentioned before, we perform PCA to identify the system model by separating the data into two subspaces: principal component subspace (PCS) and residual subspace or null space (RS). The selection of *k* (the number of PCs that should be retained in PCS) is critical to get the right model order in RS, that is, *n-k*. Improper choice of number *k* leads to incorrect partitions between these two spaces. For instance, retaining too few principal components does not capture the total informative variance of the variables, and, at the mean time, a portion of the system variability

will be introduced to the RS, which generally reduces the sensitivity in fault detection and the resolution in isolation. By appropriately determining the number of principal loading vectors (PCs) k, the true-signal variation can be decoupled from the noise variation, and the two types of variation can be monitored separately. This is not easy, but a number of heuristic rules have been proposed to improve the results.

80% Variance Test (Malinowski, 1991)

Determines the number of PCs by counting PCs until the cumulative variance explains at least 80% of the total variance.

Scree Test

In the scree plot, the plot of variance captured by each PC versus the sequence number of the PC, the dimension of the PCS is determined by locating the eigenvalue λ, which is where the profile shows an elbow. The identification of this elbow can be ambiguous and difficult to automate in implementation.

Eigenvalue-one rule

In the case of standardized data, retain those PCs whose eigenvalues are greater than one. This makes sense because SNR is commonly greater than one.

Parallel Analysis

Compares the variance profile of latent variables to that obtained by assuming independent observation variables. k is determined at the point where the two profiles cross. This approach ensures that significant correlations are captured in the process space, i.e., PCS. This method is attractive since it is intuitive and easy to implement. Several researchers, reporting on practical applications of this method, have reported that it is very effective (Wenfu Ku, *et al.*, 1995).

Cross-validation (Wold, 1978)

This method uses Predictive Sum of Squares (PRESS) statistics to determine the optimum number of the PCs,

$$PRESS(k) = \frac{1}{nN_i}\left\|\mathbf{Y} - \hat{\mathbf{X}}\right\|^2$$

Here k is the number of loading vectors retained to calculate $\hat{\mathbf{X}}$, i.e., the dimension of PCS. In cross-validation, we follow the steps:

(1) The training data set is divided into several groups (e.g., 3 ~5 groups chosen orderly or randomly). (2) The PRESS(k) for each group "i" is computed based on the PCA model built for the data in all the other groups. (3) Repeat the first two steps and plot the summation of PRESS statistics vs. k. The dimensionality is determined by locating the minimum of the plot.

AIC (Akaike Information Criterion, Akaike, 1974)

AIC, often used in system identification in dynamic cases, determines the model order k by minimizing an information theoretic function of k. For PCA modeling, under the assumption of Gaussian distribution, AIC(k) is defined as:

$$\mathrm{AIC}(k) = N \ \ln\left|\mathbf{A}_k^T \mathbf{A}_k \mathbf{Y}\mathbf{Y}^T \mathbf{A}_k^T \mathbf{A}_k\right| + 2k$$

where N is the number of samples, and $\hat{\mathbf{E}}\hat{\mathbf{E}}^T = \mathbf{A}_k^T \mathbf{A}_k \mathbf{Y}\mathbf{Y}^T \mathbf{A}_k^T \mathbf{A}_k$ is the estimated variance of the white noise error (i.e., the prediction error), a decreasing function of k. The term $2k$ is a "penalty" for over-fitting.

Minka (2000) introduced the Laplace evidence method to determine the PCA model order, but this method requires the normality assumption for the signal. Other sophisticated statistical criteria have been reported in the last decade (see Nounou and Bakshi, 2002; Everson, 2000; Bishop, 1998; Rajan, 1997).

Section 2.5.3 will present a new and meaningful method for model order determination.

2.4 Scaling of Data Prior to PCA

2.4.1 Review of Scaling Methods

In doing principal component analysis on a data set, we assume that all our data are on a comparable scale. If this is not the case, then certain elements of the data set have to be adjusted in order that misleading dominance does not occur. Scaling of data changes the covariance matrix and consequently affects the principal components. If no scaling is employed (without zero-mean), the resultant matrix will be a product of the second moment matrix. It will reduce to a covariance matrix if the mean is subtracted and will reduce to a correlation matrix if the data is also scaled to have unit variance. Scaling is meaningful with respect to variance adjustment and mean adjustment.

- As for the variance aspect, the original variables are in different units. In this case, the operations involving the trace of the covariance matrix have no meaning. For instance, if a variable is measured in centimeters, its variance is 10,000 times what it would be if it were measured in meters. This variable would then exert considerably more influence on the shaping of the PC's since PCA is concerned with explaining the maximum variability (equation 2.2a, 2.2b).

- As for the data mean, the second moment $\mathbf{Y}\mathbf{Y}^T$ may differ from the covariance matrix widely, and give undue weight to certain variables.

When the units are different, a typical type of scaling is to make variances the same (i.e., standard units), which gives a correlation matrix. Often some variance-stabilizing transformation such as, for example, log transformation, is done. But a widely accepted scaling method is to convert the variables to zero mean and unit variance, i.e., auto-scaling. Sumpter (1997) has provided an illustrative example of how scaling factors can change the shape of eigenvalue plots. The various data scaling methods are summarized below:

Unit-variance scaling

As mentioned before, unit-variance scaling is used to standardize the variance:

$$y_{j,s} = y_j / \sigma_j$$

Auto-scaling

The same as #1 above, but zero-mean the data.

$$y_{j,s} = \left(y_j - \bar{y}_j\right) / \sigma_j$$

Scaling by error covariance matrix Σ_e

When we talk about scaling, we commonly consider a diagonal matrix **D** so that $\mathbf{Y}_s = \mathbf{DY}$ is scaled data. But here Σ_e is a symmetric matrix that may contain non-diagonal elements. If this matrix is available and has full rank, then scaling the data as $\mathbf{Y}_s = \Sigma_e^{-1/2}\mathbf{Y}$ gives the optimal estimate of **X** using PCA. This method is discussed later.

Scaling by domain

All data are scaled to range [0 1]. For data $\mathbf{Y}_{n\times N}$ in equation 2.1:

$$\left(y_{i,j}\right)_s = \left| \frac{y_{i,j}}{\max_k \left(y_{i,k}\right)} \right| \qquad \forall_{k,j=1\cdots N} \;\&\; \forall_{i=1\cdots n}$$

Scaling by range

The median is used instead of the mean, and upper or lower 75% (or any other alternative) values are used instead of standard deviation in auto-scaling.

$$\left(y_i\right)_s = \frac{y_i - \text{median}(y_i)}{y_{i(+75\%)} - y_{i(-75\%)}}$$

This scaling method is robust to outliers.

Non-linear scaling methods

There are different kinds of non-linear scaling methods, such as Kernel mapping, etc. These methods are designed to cope with the non-linear relationships in variables.

2.4.2 Effects of Scaling on PCA Modeling

A linear transformation using a nonsingular matrix **D** can be generally defined as

$$\mathbf{Y}_s = \mathbf{DY} = \mathbf{DX} + \mathbf{DE}$$

where **D** is of full rank. If **D** is a diagonal matrix, then it simply scales the data. However, scaling may change results of PCA modeling. For instance, applying PCA on a zero-mean data set will result in different outputs from the case of applying PCA on an auto-scaled data set. In other words, there is no one-to-one correspondence between the PC's obtained from a correlation matrix (if auto-scaling applied) and those obtained from a covariance matrix (Jackson, 1991).

In following, we will mainly discuss the effects of a linear scaling method on PCA modeling.

Case of noise-free data and D is diagonal

Since **D** is nonsingular, the rank of $\mathbf{Y}_s\mathbf{Y}_s^T$ is same as the rank of $\mathbf{YY}^T$. If we apply PCA to the scaled sample covariance matrix $\mathbf{Y}_s\mathbf{Y}_s^T/(N-1)$, the transpose of the last k orthonormal eigenvectors corresponding to the zero eigenvalues represent a basis for the residual space (RS) which is orthogonal to the scaled data vectors $\mathbf{Y}_s$. If we denote this basis for RS by a k dimensional linear model $\mathbf{A}_s$, then we have:

$$\mathbf{A}_s\mathbf{Y}_s = \mathbf{0}$$

Using the definition of scaling we have:

$$\mathbf{A}_s\mathbf{DY} = \mathbf{0} \qquad \ldots\ldots\ldots(2.19)$$

From equation 2.19 we see that the rows of the matrix $\mathbf{A}_s$ is also a basis for RS. Thus, in the absence of measurement errors we identify an exact basis for the model, i.e., the basis for RS even if we apply PCA to scaled data $\mathbf{Y}_s$.

Case when noise E is a constant offset matrix

In this case we can reduce the rank of Σ_Y to be the same as Σ_X as long as we perform the zero-mean operation on the training data. See figure 2.2, where the error **E** has been totally removed and the problem reduces to the above discussed noise-free data case.

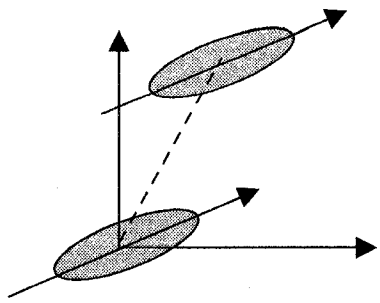


Figure 2.2 Constant offset does not affect the PCA results

Case when noise E is a random noise matrix

If **E** is not an offset matrix but a random noise matrix so that Σ_e is of full rank, we cannot simply apply zero-mean operation to find the rank of the signal.

For scaled data

$$\mathbf{Y}_s = \mathbf{DY} = \mathbf{DX} + \mathbf{DE}$$

the population covariance matrix is:

$$\Sigma_{Y_s} = \mathbf{D}\Sigma_Y\mathbf{D}^T = \mathbf{D}\mathbf{S}_X\mathbf{D}^T + \mathbf{D}\Sigma_e\mathbf{D}^T$$

The eigenvectors of Σ_Y and Σ_{Ys} do not bear any simple relation to each other (Morrison, 1967).

D is a linear transformation of data **Y** that shrinks or stretches the data and, consequently, usually rotate the eigenvectors in the data space. It makes no difference in the results of PCA analysis as long as we look at the transposed space spanned by **all** the latent variables for interpreting the data without any loss of information. However, we always expect fewer dominant latent variables (or PCs) to count for the system variability. In this context, care should be paid on the effects of scaling. This is because **D** may change the sequence of PCs and twist the PCs' directions if unequal levels and/or correlated noise exist. Improper scaling may assign PCs to residual space (RS) and move latent variables in RS to the principal component subspace (PCS). Hence, poor scaling fails to distinguish between the system variability and random-noise variability.

If we assume that all true-signal variances are much larger than all noise variances, then, the effect of scaling is not significant: without scaling, the orthogonal eigenvectors of Σ_Y corresponding to m small eigenvalues can be used as an estimate for model **A** and therefore we can skip the scaling of the data. However, if the variance of at least one variable in **X** is less than

the noise variance in another variable, but SNRs are quite large, then unit-variance scaling will be helpful. For example, we have

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

where std(y_1) = 0.5 < std(e_2) = 0.8, but std(e_1) = 0.05, std(y_2) = 8. In this case, the unit-variance scaling will help signals dominate the noise. Here, scaling does improve the quality of the identified model, but the effect is hard to quantify.

The following example shows that the quality of PCA model results from the choice of scaling.

An example - Choice of scaling methods yields different PCA-modeling results

This simple example is used to illustrate the effect of different choice of scaling the data on PCA results. This example system consists of two variables x_1 and x_2, which are equal to each other. Measurements of the two variables are generated with very different signal-to-noise ratios, i.e.,

$$SNR\_x_2 = 10 \times SNR\_x_1$$

where SNR is simply defined as the ratio of standard deviation of the true signal and the measurement noise (and/or any unaccountable deviations).

Figure 2.3 compares the results of applying PCA to the generate data for the following two different data scaling choices:

(1) Auto-scaling.
(2) Scaling by error covariance matrix Σ_e.

The principal vectors **PC1** and **PC2** are calculated from case (1) (the red arrows) and from case (2) (the bold black arrows) respectively. We can see that the first PC from case (2) aligns with the fine dotted line, the true relationship "curve" of the data, which passes through the middle of the data cluster. However, the traditional PCA approach, the case (1), gives a misleading model direction that deviates from the true model "curve". In generating this figure, we can also see that we need a higher signal-to-noise ratio in x_1 to obtain a similarly good fit in case (1) versus case (2). This is because high SNR value makes the data cluster to be in a narrowed envelope and therefore makes the red arrows close to black arrows. Without high SNR, auto-scaled PCA cannot provide a perfect fit even when the sample size is infinitely large. In other words, auto-scaled PCA gives a biased estimate of process model.

In fact, case (2) is the optimal scaling method we can choose of all the alternatives. The verification of this point will be given in the following section.

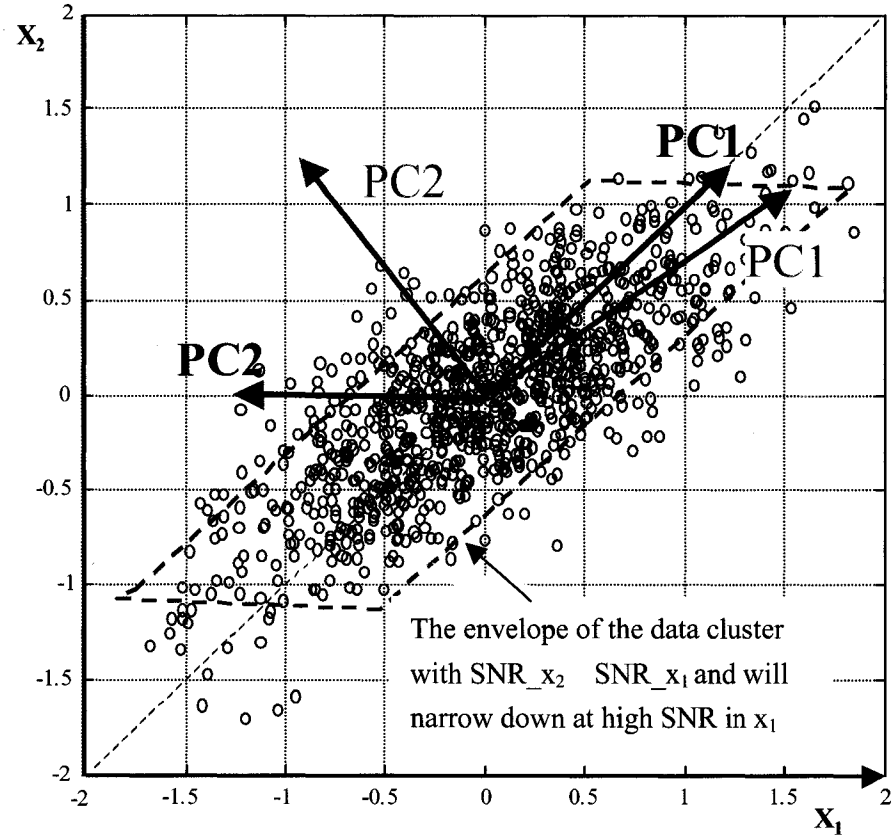


Figure 2.3 Different scaling methods result in different PCA modeling results

2.5 The Optimal Solution for PCA Modeling – OSPCA and IPCA

From the example shown in figure 2.3, we realize that the direction of the maximum variability in the data does not necessarily mean the direction of the underlying model. In model identification or extraction, our objective is to find a sound PCA-based approach to an optimal estimate of principal component space PCS or its complementary part, the residual space RS (In FDI problem, the later is critical). Pursuing this objective, we should look at the PCA modeling problem carefully.

2.5.1 Objectives

Let us go back to the very beginning of linear model identification. Given any observation of an $n\times1$ data set:

$$\mathbf{y} = \mathbf{x} + \mathbf{e}$$

where $\mathbf{x}$ is the true signal and $\mathbf{e}$, the random noise in the observation.

In general, signals can have *spatial correlation* and/or *temporal correlation* (refer to 6.2.1). The true signal $\mathbf{x}$ can be *stochastic* or *deterministic* with spatial covariance or second moment $\mathbf{S}_x$ (rank = k) and temporal covariance Φ_X. The measurement errors $\mathbf{e}$ is multivariate with spatial covariance Σ_e and temporal covariance Φ_e. If $\mathbf{e}$ is iid-normal, the temporal covariance Φ_e will have identical diagonal elements. If the data matrix consists of the measured data vectors at each time, then the PCA model determines the spatial relationships among different variables, i.e., the *spatial correlation* in $\mathbf{x}$. If the data matrix is augmented with time-lagged variables, then it is possible for PCA to identify both the temporal and spatial relationships that exist among the variables. This is referred to as dynamic PCA. This thesis is concerned with identifying the steady state spatial relationships among the variables.

The *steady state* PCA-based modeling objectives can be stated as follows:

(a) To decouple the signal variability from the noise variability, and determine the number of PCs.
(b) To decouple the spatial covariance $\mathbf{S}_X$ from Σ_e.
(c) From data correlation structure obtained after steps (a) and (b), estimate the underlying model using an appropriate approach such as PCA.

From these points of view, **PCA is a method for decoupling signals and noises.**

If part of the noise **E** has more power than the excitation in true signal **X**, i.e., SNR<1, task (a) becomes difficult. In doing regular PCA in the time domain, there is no sense in discriminating noise from true signals if SNR<1 in some variables. What we get in PCA is either a signal model or a noise model or a mixture of the two. We can obtain some hints from checking the distribution patterns of scores in PCS or RS if we have enough data. We can also apply filters (via wavelet functions or EWMA), for instance, one can use MSPCA in the frequency domain to estimate a model in a predefined band of frequency. This, however, is beyond the scope of this thesis. In any case, for process modeling, we can always expect an agreeable data quality with SNR>1.

Although (b) helps get a good result in step (a), step (b) itself is the critical one in model identification requiring specific techniques: statistically-based scaling (described here) or maybe some direct statistical approaches if analytically or numerically applicable. Step (c) performs what ordinary PCA does.

As a direct statistical approach for step (b), Wentzell *et al.* (1997) proposed a MLE-based iterative solution for PCA modeling, when both the spatial and temporal correlations of the measurement errors are completely known. Tipping and Bishop (1999a, 1999b) have given a solution for doing probabilistic PCA under tight limitations. Instead of maximizing variability, they propose an algorithm called probabilistic PCA to simultaneously estimate the maximum likelihood eigenvectors and the maximum likelihood noise variance respectively. We can see from the algorithm, although the paper does not mention this, that probabilistic PCA is implicitly more tolerant to low SNR than ordinary PCA. However, Tipping and Bishop restrict the noise variance to be diagonal with identical variances for different variables. Notice that the assumptions for probabilistic PCA are the same as the aforementioned implicit assumption for ordinary PCA (LS based), with the added assumption of normality for the excitation of the true signal **X**. This assumption represents the cost of providing analytical solutions for estimating mean, model and error covariance in one shot. Minka (2000) subsequently proposed the sophisticated Bayesian method for determining the subspace order.

Notice that, when we come to look at a data set, the ranks and structures of the underlying $\mathbf{S}_X$ and Σ_e in the data are generally different, which makes PCA modeling difficult.

To follow the aforementioned statistical approaches we have the following challenges:

- Generally Σ_e is unknown.
- The true signal **X** rarely follows multivariate normal distribution.

In this section, we will describe an optimal scaling method that gives PCA modeling a strong theoretical basis when measurement errors in different variables are unequal and correlated. A procedure for simultaneously estimating both the model and error covariance matrix using an

iterative PCA technique, which has been recently developed by Narasimhan and Shah (2004), is also described, since this method was proposed as the solution for the optimal scaling PCA problem raised at early phase of this thesis study. This method, therefore, is consequently used as a starting point of the thesis.

2.5.2 Optimally Scaled PCA (OSPCA) with Known Error Covariance

Problem Formulation

Assume the error covariance matrix e is known as Σ_e. If Σ_e has full rank, $\Sigma_e = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is the square root of Σ_e. Let scaling matrix $\mathbf{D} = \mathbf{L}^{-1} = \Sigma_e^{-1/2}$, then the scaled data is:

$$\begin{aligned} \mathbf{Y}_s = \mathbf{D}\mathbf{Y} &= \mathbf{L}^{-1}\mathbf{X} + \mathbf{L}^{-1}\mathbf{E} \\ &= \mathbf{X}_s + \mathbf{L}^{-1}\mathbf{E} \end{aligned} \qquad \text{.....(2.20)}$$

If Σ_e is diagonal, this is equivalent to scaling $\mathbf{Y}$ by the standard deviation of the measurement errors. From equation (2.20) we have

$$E\left\{\mathbf{Y}_s\mathbf{Y}_s^T\right\} = E\left\{(\Sigma_e)^{-1/2}\mathbf{X}\mathbf{X}^T(\Sigma_e)^{-1/2}\right\} + E\left\{(\Sigma_e)^{-1/2}\mathbf{E}\mathbf{E}^T(\Sigma_e)^{-1/2}\right\}$$

and

$$\Sigma_{Y_s} = \mathrm{S}_{X_s} + \mathbf{I}$$

From the *Eigenvalue Shift Theorem*, it is known that if $\lambda, \mathbf{v}$ is an eigenpair of a matrix $\mathbf{M}$, and α is any constant, then $\lambda - \alpha, \mathbf{v}$ is an eigenpair of the matrix $\mathbf{M} - \alpha\mathbf{I}$.

Then the following two important properties can be easily derived:

- The eigenvectors of Σ_{Y_s} are identical to those of S_{X_s}.
- The eigenvalues of Σ_{Y_s} are equal to those of S_{X_s} shifted by unity.

The first means that we can obtain the eigenvectors of S_{X_s} immediately by using PCA on the scaled data $\mathbf{Y}_s$, and therefore model $\mathbf{A}$ for original data $\mathbf{Y}$ can be obtained from $\mathbf{A}_s$ in equation 2.19. The second means that the smallest eigenvalues corresponding to model $\mathbf{A}_s$ are unity.

In fact, analogous to the discussion in the PCA section, the optimal scaling method we have described here is equivalent to the formulation given below:

For example, suppose we have data matrix $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. If we weight the data properly and then perform PCA, the formulation is given by:

$$\min_{\mathbf{V}} J_{\text{IPCA}} = \sum_{i=1}^{N} \left(\mathbf{V}^T y_i\right)^T \left(\mathbf{V}^T \Sigma_e \mathbf{V}\right)^{-1} \left(\mathbf{V}^T y_i\right) \qquad \text{.....(2.21)}$$

$$s.t. \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}$$

to find orthogonal directions **V** so that the projections of **Y** represent maximum variability in the data. Notice that if data **Y** is scaled by $\mathbf{L}^{-1}$, then $\Sigma_{e,s} = \mathbf{I}$, equation 2.21 will be equivalent to PCA (see equation 2.13).

The Justification of OSPCA - Equivalency of OSPCA and MLPCA

Why we call the method as Optimal Scaled PCA? The following addresses the verification of this method by comparing it with a reported MLE approach for PCA modeling.

Wentzell *et al.* (1997) give a MLE approach for model ID using PCA (MLPCA) when Σ_e is known. This method is implemented by an iterative procedure, instead of scaling the data, until the MLEs of observations **Y** are obtained.

MLPCA estimates the model that maximizes the likelihood of estimating the true principal component scores and loading vectors given the measured variables, or, equivalently, maximizes the probability density function of the measurements by selecting the noise-free principal component scores, loading vectors, and the true rank of the data matrix "*k*," as

$$\left\{\hat{\mathbf{P}}, \hat{\mathbf{T}}\right\}_{PCA} = \arg\max_{\hat{\mathbf{P}}, \hat{\mathbf{T}}} \left(\mathbf{Y} | \hat{\mathbf{P}}, \hat{\mathbf{T}}, k\right)$$

$$s.t. \quad \hat{\mathbf{T}} = \hat{\mathbf{P}}^T\mathbf{Y}, \text{ and } \hat{\mathbf{P}}^T\hat{\mathbf{P}} = \mathbf{I}$$

Assuming the error **E** is normally distributed, the above equation can be reduced to minimizing the sum of square errors normalized by Σ_e (Nounou, 2002):

$$\left\{\hat{\mathbf{P}}, \hat{t}_j\right\}_{PCA} = \underset{\hat{\mathbf{P}}, \hat{t}_j}{\text{argmin}} \sum_{j=1}^{N} \left(y_j - \hat{x}_j\right)^T \Sigma_e^{-1} \left(y_j - \hat{x}_j\right) \qquad \text{.....(2.22a)}$$

$$s.t. \quad \hat{x}_j = \hat{\mathbf{P}}\hat{t}_j, \text{ and } \hat{\mathbf{P}}^T\hat{\mathbf{P}} = \mathbf{I} \qquad \text{.....(2.22b)}$$

where the *principal component score* is solved from the *data reconciliation* problem as

$$\left\{\hat{t}_j\right\}_{DR} = \left(\hat{\mathbf{P}}^T\Sigma_e^{-1}\hat{\mathbf{P}}\right)^{-1} \hat{\mathbf{P}}^T\Sigma_e^{-1} y_j \qquad \text{.....(2.23)}$$

A brief discussion of the data reconciliation problem will be given in Chapter 4. If we substitute $y_{j,s} = \mathbf{L}^{-1} y_j, \hat{x}_{j,s} = \mathbf{L}^{-1}\hat{x}_j\ \hat{t}_{j,s} = \hat{t}_j, \hat{\mathbf{P}}_s = \mathbf{L}^{-1}\hat{\mathbf{P}}$ into equation 2.22a, 2.22b, and 2.23 but change the constraint $\hat{\mathbf{P}}^T\hat{\mathbf{P}} = \mathbf{I}$ to $\hat{\mathbf{P}}^T\left(\mathbf{L}\mathbf{L}^T\right)^{-1}\hat{\mathbf{P}} = \mathbf{I}$, then, knowing that $\mathbf{L}^{-1} = \Sigma_e^{-1/2}$, we can find the formulation of MLPCA (equation 2.22a, 2.22b and 2.23) is reducible to the formulation of PCA (equation 2.6). So the MLPCA approach is equivalent to the OSPCA approach presented above. The only difference is that in MLPCA the estimated loading matrix **P** is orthogonal and the estimated score matrix **T** is not orthogonal. Alternatively, in optimally scaled PCA, both loading matrix and score matrix are orthogonal in the scaled domain. In the un-scaled domain, it turns out that in optimally scaled PCA, the estimated loading matrix **P** is no longer orthogonal but the estimated score matrix **T** is orthogonal.

2.5.3 Iterative Principal Component Analysis (IPCA) with Unknown Error Covariance

Estimating Error Covariance and the Model Simultaneously

In the chemical industry, it is quite common that we do not know the error covariance matrix. It is an advantage if we can estimate error covariance because we can then apply the above-mentioned optimal scaling approach to get a model based on noisy data.

Narasimhan and Shah (2004) proposed a method to apply the OSPCA in an iterative loop to estimate error covariance matrix Σ_e and model **A** simultaneously. In doing this, model **A** is initially estimated by PCA with "proper" scaling, using the initial value of Σ_e. Then model **A** and error covariance Σ_e are iteratively updated in an alternative way until the convergence criteria has been met. This method was referred to as *Iterative PCA* (IPCA). The basic assumptions and details of this algorithm are described below.

Basic Assumptions for IPCA

As outlined in the following sections, which give the implementation of the IPCA algorithm, IPCA requires limited assumptions:

Given data $\mathbf{Y}_{(n\times N)} = \mathbf{X}+\mathbf{E}$, where n = number of variables, and N = number of observations, the following assumptions are required:

- Relationships between variables can be captured by a linear model: $\mathbf{AX} = \mathbf{0}$.
- Measurement error **E** is iid-normal and independent of **X**.
- The errors cannot be significantly dependent on each other so that Σ_e has full rank, but measurement errors are allowed to be unequal and correlated to some extent.
- There is sufficient information for **X** at least for the training data set, e.g., SNR > 1.

Algorithm of Iterative PCA (IPCA)

1) Given an initial estimate $\hat{\mathbf{A}}(0)$ by using PCA, we can obtain an estimate of $\hat{\Sigma}_e$ using equation 2.24:

$$\mathbf{r}(j) = \hat{\mathbf{A}}(0)\mathbf{y}(j) \sim N\left(0, \hat{\Sigma}_r\right) \quad \Sigma_r = \hat{\mathbf{A}}(0)\hat{\Sigma}_e\left(\hat{\mathbf{A}}(0)\right)^T$$

$$\min_{\hat{\Sigma}_e} N\log\left|\hat{\mathbf{A}}(0)\hat{\Sigma}_e\hat{\mathbf{A}}(0)^T\right| + \sum_{j=1}^{N}\mathbf{r}_j^T(\hat{\mathbf{A}}(0)\hat{\Sigma}_e\hat{\mathbf{A}}(0)^T)^{-1}\mathbf{r}_j \qquad \ldots\ldots\ldots(2.24)$$

2) Left scale the data $\mathbf{Y}$ using $\mathbf{D} = \left(\hat{\Sigma}_e(k)\right)^{-1/2}$, perform PCA to get an updated estimate of model $\hat{\mathbf{A}}(k+1)$, then repeat (1) and (2) iteratively until convergence. The convergence criteria used is to check that the singular values have not changed significantly.

Note that the algorithm of IPCA gives a MLE applied in residual space to estimate the error covariance matrix $\hat{\Sigma}_e$ given the model $\hat{\mathbf{A}}$ is true. In fact, as shown later, this algorithm gives the correct model order using the eigenvalue equal-to-one rule.

Limitation of IPCA

Depending on the number of constraints and number of variables, it may not be possible to solve for all elements of Σ_e by above-mentioned algorithm. Typically, the number of spatial relations m is less than the number of variables n. In such cases, if we attempt to estimate all diagonal and off-diagonal elements of Σ_e, there is no unique solution that satisfies equation 2.24. One possibility is to assume that Σ_e is diagonal, which is often true in chemical engineering, and estimate only the n diagonal elements of Σ_e. Even in this case a non-degenerate estimate for Σ_e can be obtained only when:

$$\frac{m(m+1)}{2} \geq n \qquad \ldots\ldots\ldots(2.25)$$

We can impose lower bounds on the elements of Σ_e: limits of the measurement accuracy of instruments; we can also define the upper bounds as the observed variations in the variables. For a special case, in which we know that a few off-diagonal elements in Σ_e is non-zero and to be estimated, we can use this constructional knowledge to help solve the problem. For a diagonal Σ_e, condition (2.25) is generally satisfied for most processes.

The IPCA approach thus provides an estimate of the process constraint model even when the errors in different variables do not have identical variances. Furthermore, an estimate of the error covariance matrix is also obtained simultaneously along with the model. These two pieces of information extracted from the data are vital to applying data reconciliation and gross error detection strategies. In the following two chapters, the efficacy of the IPCA method to extract the process model accurately is studied. Using the estimated model and error covariance matrix, the

well known techniques of data reconciliation and gross error detection are applied to demonstrate that more accurate state estimates and improved isolation of sensor faults can be obtained than what can be achieved using PCA and contribution plots.

2.6 Concluding Remarks

The discussion in this chapter began by reviewing the PCA method as a tool for model identification. Three formulations are given for the PCA method from different points of view: maximizing the variability in the data to find PCS, minimizing residuals in the residual space (Constraint Model ID), and Total Least Squares Regression. To help understand PCA, the comparison between PCA and the well-known LS-Regression was given. Then an optimum solution for doing PCA – the optimally scaled PCA (OSPCA) is described to decouple the covariance (or the second moment) $\mathbf{S}_X$ of true signal from the covariance Σ_e of measurement noise. A comparison between the proposed OSPCA and Maximum Likelihood PCA (MLPCA) showed the equivalency of the two methods. Finally improved method, IPCA, is described and the advantages of this method were discussed. In addition, various methods for data scaling and model order determination were discussed, where necessary, to help understand the problem. The contributions in this chapter can be summarized as follows:

- PCA analysis is an approach for model identification.
- PCA can be useful when there is a severe high-degree of correlation present in the data set.
- PCA works as a Total Least Squares (TLS) regression and gives a maximum likelihood estimate only under the assumption that all measurement error variances are equal.
- The PCA method and Least Squares (LS) regression method are similar, with each having advantages and disadvantages.
- Auto-scaled PCA gives a biased estimate of process model, and the heuristic rules for model order determination often give an ambiguous answer.
- If the error covariance is known, optimal scaling method gives an ML estimate with PCA and is equivalent to the iterative MLPCA method that is proposed in the literature.
- IPCA estimates the error covariance and the model in the sense of MLE. Although it is not exactly MLE, it is an optimum solution based on the information that we have. At the same time, IPCA gives a theoretically-based rule for model order determination.
- Usually not all elements of the error covariance can be successfully estimated by IPCA, due to limited number of constraints. But *a priori* knowledge of the error covariance structure is a great help.

Chapter 3

Steady State Model Identification

3.1 Introduction

In this chapter, the ability of IPCA to obtain accurate process models is studied through simulation and the results compared with those obtained using PCA. The accuracy of state estimates obtained using IPCA and PCA is also compared.

3.2 Case Study

3.2.1 The Flow Network Example

We first describe the flow network which is used for all the simulation studies reported in this thesis. Figure 3.1 shows a schematic of a flow network where 15 flow rates are read from 15 flow-rate sensors in the process. Given a set of observations for these 15 flow rates, the natural questions that arise are:

- Can all the data be reconciled?
- If not, how can we detect the existence of at least one sensor fault in the observed data and the time instant when it occurs?
- Which sensor is responsible for the fault?
- By how much and in which direction should we recalibrate the sensor?
- Is any sensor fault easier to detect over others and can we obtain a measure of detectability?

It is not possible to answer these questions by examining Figure 3.1. To answer these questions we apply statistical techniques and algebraic operations with the understanding that the knowledge of model structure and measurement error structure is helpful.

For this problem, the model, i.e., the group of mass balance equations, is readily available directly from the flow network. However, we are interested in more general cases where limited information is available about the model structure and measurement error covariance matrix are identified from operating data. An appropriate approach will be suggested to identify the model. Detailed discussions of this example will be covered in the forthcoming chapters of the thesis.

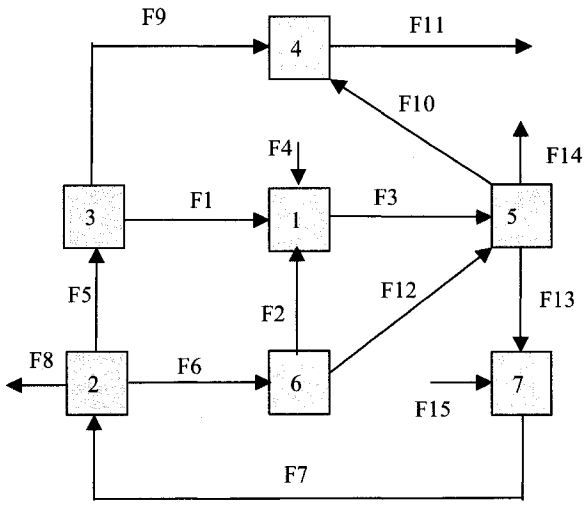


Figure 3.1 Flow network schematic

3.2.2 Data Generation for the Flow Network Example

Fault-Free Data

In the flow network example, the true (error-free) flow rates of the 15 variables are first generated as follows. Since there are 7 independent steady state flow balance constraints relating the 15 variables, the process has 8 degrees of freedom. Flow variables F1, F2... are chosen as the independent variables and their true values at each time instant are generated, and the true values of the remaining variables are calculated so that they satisfy the mass balances. The true values of the eight independent flows are generated by passing white noise through 8 different first order filters. A data set consisting of 200,000 samples is generated. Part of the data set also consists of a combination of different forms of signals: constant mean, square impulse, sinusoidal and random binary inputs.

After the error-free data set is generated, measurement errors are added to the true values at each time instant to simulate the measured values. The measurement error at each time instant is a random normally distributed vector with a 0-mean and a specified standard deviation (noise level). Changing these noise levels means changing the measurement accuracy and, consequently, changing the Signal-to-Noise Ratio (SNR).

The data set of 200,000 samples, is divided into 100 segments $\mathbf{Y}_1 \sim \mathbf{Y}_{100}$, as shown in Figure 3.2, Each segment $\mathbf{Y}_k = [y_1, y_2, \ldots, y_i, \ldots, y_{2000}]$ contains 2000 observations. PCA and IPCA are applied to each of these segments to identify the steady state model.

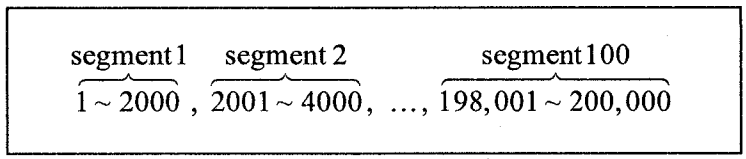


Figure 3.2 100 segments of fault-free data for PCA modeling

3.3 Comparative Study of PCA and IPCA for Model Identification

3.3.1 Performance Criteria

To evaluate the accuracy of the model identified using PCA or IPCA, we need to quantify the mismatch between the identified constraint matrix and the true constraint matrix for this process. The model accuracies are compared in terms of (i) the angle between two subspaces defined by the row spaces of the linear constraint model and (ii) the total absolute error reduction achieved by using the models.

Angle between subspaces

Given two subspaces $\mathbf{A}$ ($n \times m_1$) and $\hat{\mathbf{A}}$ ($n \times m_2$), if a and â are two arbitrary vectors in these two subspaces respectively, the angle (the difference) between the two subspaces is defined as:

$$\theta\left(\hat{\mathbf{A}}, \mathbf{A}\right) \equiv \max_{\substack{\hat{\mathbf{a}} \in \hat{\mathbf{A}} \\ \hat{\mathbf{a}} \neq 0}} \min_{\substack{\mathbf{a} \in \mathbf{A} \\ \mathbf{a} \neq 0}} \theta(\hat{\mathbf{a}}, \mathbf{a})$$

If $\mathbf{A} = \mathbf{QR}, \hat{\mathbf{A}} = \hat{\mathbf{Q}}\hat{\mathbf{R}}$ and $\mathbf{Q}$, $\hat{\mathbf{Q}}$ are orthonormal matrices, this definition is equivalent to:

$$\cos^2\theta = \min_{\mathbf{x}} \left\| \left(\hat{\mathbf{Q}}\mathbf{x}\right)^T \mathbf{Q} \right\|_2$$
$$s.t. \left\| \hat{\mathbf{Q}}\mathbf{x} \right\| = 1,$$
$$\mathbf{x} \neq \mathbf{0}$$

From this definition, we know that if $\hat{\mathbf{A}}$ is a similar linear transformation of $\mathbf{A}$, $\theta = 0$, *i.e.*, there is no model mismatch in the estimation. In MATLAB, we use subspace ($\mathbf{A}$, $\hat{\mathbf{A}}$) to calculate theta. The algorithm for this command performs Q-R decomposition of $\mathbf{A}$ and $\hat{\mathbf{A}}$, followed by singular value decomposition of matrix $\mathbf{Q}^T\hat{\mathbf{Q}}$. We can see why this is by looking at the Lagrangian:

$$\Re = \left(\hat{\mathbf{Q}}\mathbf{x}\right)^T \mathbf{Q}\left[\left(\hat{\mathbf{Q}}\mathbf{x}\right)^T \mathbf{Q}\right]^T - \lambda\left(\left(\hat{\mathbf{Q}}\mathbf{x}\right)^T\left(\hat{\mathbf{Q}}\mathbf{x}\right) - 1\right) = 0$$

$$\frac{\partial \Re}{\partial \mathbf{x}} = 2\hat{\mathbf{Q}}^T\mathbf{Q}\mathbf{Q}^T\hat{\mathbf{Q}}\mathbf{x} - 2\lambda\hat{\mathbf{Q}}^T\hat{\mathbf{Q}}\mathbf{x} = 0$$

$$\frac{\partial \Re}{\partial \lambda} = \left(\hat{\mathbf{Q}}\mathbf{x}\right)^T\left(\hat{\mathbf{Q}}\mathbf{x}\right) - 1 = 0, \hat{\mathbf{Q}}^T\hat{\mathbf{Q}} = 1$$

$$\hat{\mathbf{Q}}^T\mathbf{Q}\mathbf{Q}^T\hat{\mathbf{Q}}\mathbf{x} = \lambda\mathbf{x}, \text{ and } \mathbf{x}^T\mathbf{x} = 1$$

so:

$$\cos\theta = \sigma_{\min}\left(\mathbf{Q}^T\hat{\mathbf{Q}}\right)$$

Reduction in TAE (Total Absolute Error)

Suppose we know x_i, the true value for an observation y_i. Then we can define the *total absolute error* (TAE) as:

$$E_1 = \sum_{i=1}^{n} |y_i - x_i|$$

If we obtained estimates of the variables from the measurements and the identified constraint model, the TAE between estimated values and observed values is given by

$$E_2 = \sum_{i=1}^{n} |\hat{x}_i - x_i|$$

where $\hat{x}_i$ is the estimate of x_i from equation 4.2, and n is the number of variables.

The reduction in TAE can be defined as:

$$\eta(\%) = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{E_1 - E_2}{E_1}\right)_j \times 100$$

The more accurate the model $\hat{\mathbf{A}}$ is, the higher is the value of η.

3.3.2 Results and Discussion

As a first comparison, we assume that the number of constraints (model order) for the process is correctly known a priori (equal to 7 for this process) and compare the models obtained using PCA and IPCA for this known model order. Figure 3.3 shows a scatter plot of the subspace angles

between the true and identified constraint row spaces for the 100 segments of training data obtained using PCA and IPCA. This figure clearly shows that the magnitudes of the angles (which represent the model estimation errors) for the models identified using PCA are consistently larger and also show greater variability than those identified using IPCA. This clearly demonstrates that the IPCA method consistently identifies a more accurate model as compared to PCA.

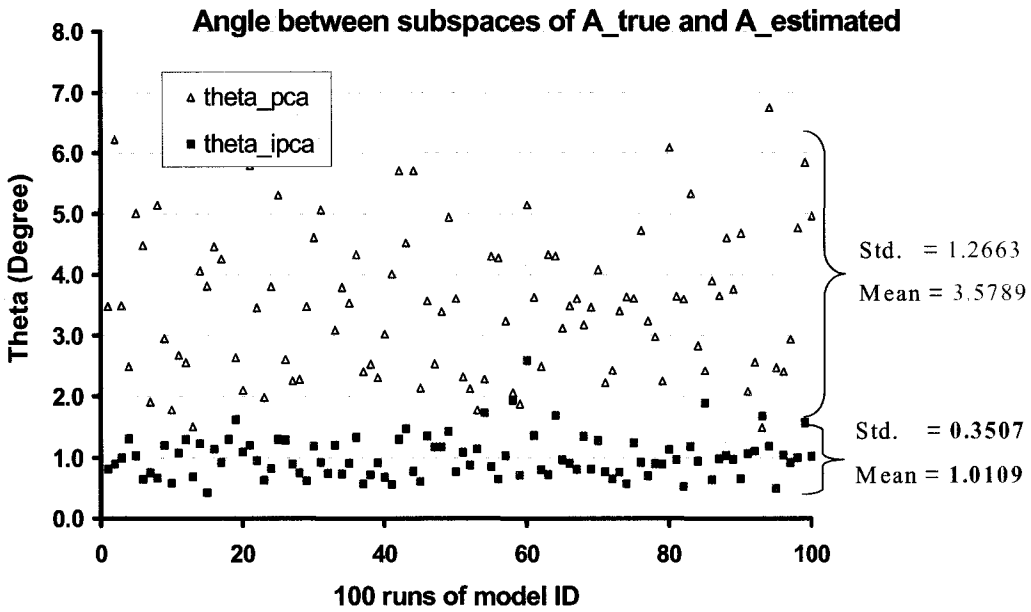


Figure 3.3 A comparison of the errors of estimated models obtained by PCA and IPCA

Since IPCA can also estimate the measurement error covariance matrix, we can compare the actual measurement error standard deviations (used in the simulation) with those estimated by the IPCA method. In this example, we assume measurement errors in different variables are independent, which implies that we are only estimating the diagonal elements of the error covariance matrix. Table 3.1 shows the actual and estimated standard deviations of measurement errors obtained using IPCA on one of the data segments. It is observed that the estimated standard deviations are quite close to the actual values, thus providing further evidence that IPCA is able to accurately estimate both the model and error covariance matrix simultaneously.

std(error)_ true	std(error)_estimated
0.8000	0.8035
0.2500	0.2404
0.5000	0.5036
0.5000	0.5045
1.5000	1.4585
0.6000	0.5942
0.1500	0.1500
0.1000	0.0812
0.4000	0.3991
0.5500	0.5587
0.0616	0.0663
0.4000	0.4069
0.0500	0.0458
0.0300	0.0281
0.0080	0.0081

Table 3.1 Estimates of standard deviations of errors

Model Order Determination

From the properties of optimal scaling PCA, we know that if we scale the data using the square root of the measurement error covariance matrix, we will as many unity eigenvalues as the model order (provided our guess of the model order is correct), while all the remaining eigenvalues will be greater than unity. In order to verify whether this criterion can be used to correctly obtain the model order using IPCA, we can examine the number of eigenvalues which are close to unity obtained for every guess of the model order and check for consistency. Figure 3.4 shows the a log plot of the eigenvalues obtained using IPCA for different guesses of the model order.

In Figure 3.4, if the model order guessed is 8 (which is an overestimate since it is one more than the actual number) then the smallest 8 eigenvalues are not all equal to unity. In fact the last eigenvalue is much less than unity as seen from Figure 3.4. This clearly implies that our guess of the model order is incorrect. On the other hand, if we guess the model order to be 7 (which is the correct value), then Figure 3.3 shows that the last seven eigenvalues are close to unity, thus confirming that our guess of the model order is correct. If we underestimate the model order (for example, we guess the model order to be 6 for this process), then we observe that the last six eigenvalues are equal to unity. However, we also note that one more eigenvalue is also close to unity suggesting that the true model order may be one more than our guess. A systematic strategy

can be devised for obtaining the precise model order using IPCA by analysing the eigenvalues as follows. We start with a guess of the model order and check if we obtain as many unity eigenvalues as our guess. If so, we increment the model order one at a time until the number of unity eigenvalues is less than the guess of the model order, at which stage we can stop. The actual model order is thus one less than the order guessed at the final stage.

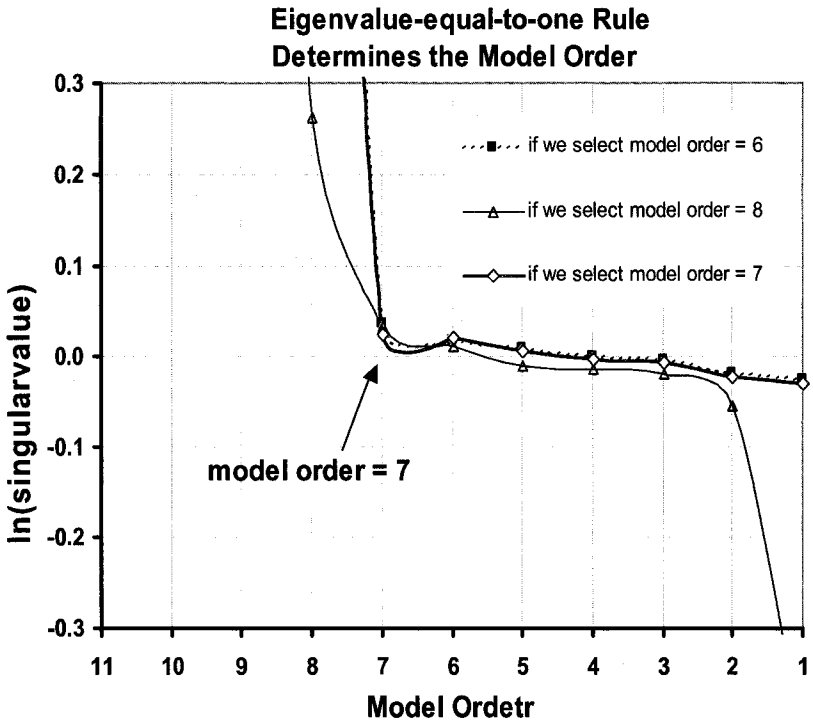


Figure 3.4. Eigenvalue-Equal-to-One Rule is used for model order determination in IPCA modeling

Model order identification using any of the heuristic rules for PCA is not so precise and can give misleading results. Figure 3.5 shows the model order obtained using PCA for the different heuristic rules (such as cumulative variance explained or scree plot). If the SNR is high, then it may be possible to estimate the correct model order as shown in Fig. 3.5a-2. But even here, there is ambiguity if we use cumulative variance explained as the heuristic for model order selection (see Fig. 3.5a-1). For low SNR values, it becomes even more difficult to accurately estimate the model order as shown by Fig. 3.5b. On the other hand, the model order can be determined precisely using the IPCA method by examining the number of unity eigenvalues obtained, regardless of whether the SNR is high or not as shown by Figs. 3.5c and 3.5d.

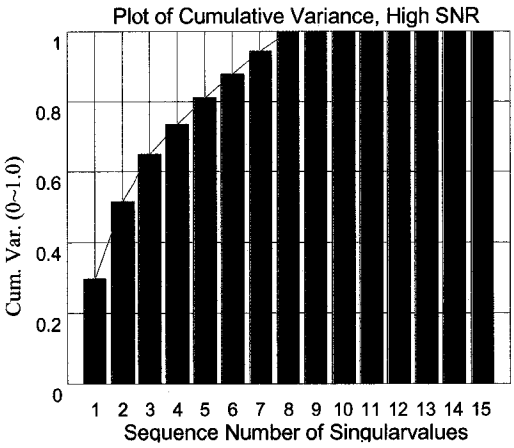
Accuracy of state estimates

Another indicator of how well IPCA is able to identify the true constraint matrix can be gauged from the accuracy of the state estimates obtained. It should be noted that the estimates of the state

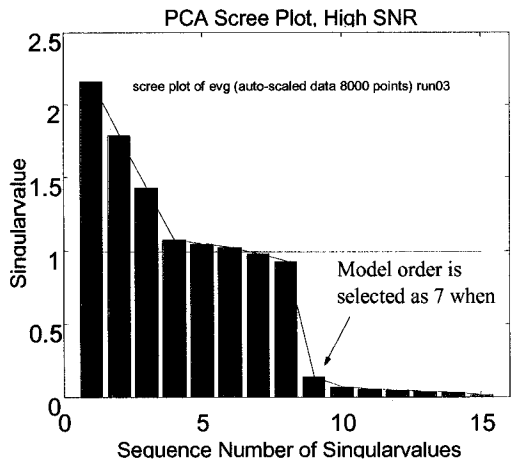
variables can be obtained using PCA from the retained principal components as given in equation 2.5. However, in the case of IPCA, while the estimates of the scaled variables $\mathbf{x}_s$ are obtained using the retained principal components as in PCA, the estimates of the original variables $\mathbf{x}$ are obtained by

$$\hat{\mathbf{x}} = \mathbf{y} - \hat{\mathbf{L}}\hat{\mathbf{L}}^T\hat{\mathbf{A}}^T\hat{\mathbf{A}}\mathbf{y}$$

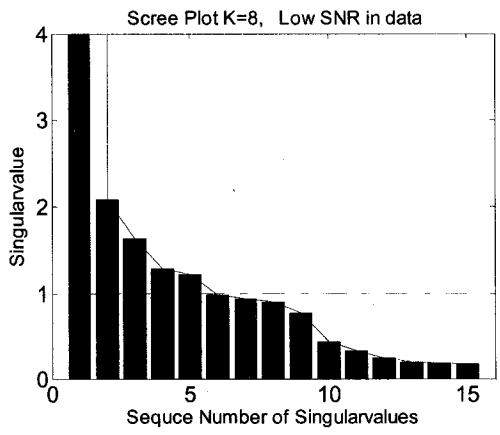
Figure 3.6, is a plot of the reduction in TAE obtained using the PCA and IPCA state estimates for corresponding to each of the 100 models. It is clear from this figure tells us that the models identified by the IPCA lead to more accurate estimates and also more consistent than those obtained using the PCA method. As a further comparison, we also plot the achievable total absolute error reduction if the true constraint matrix and true error covariance matrix are available. This is indicated by the dotted line in Fig. 3.6. It is observed that the error reduction achieved using the IPCA model and state estimates are very close to the achievable limit, thus confirming that both the model and error covariance matrix have been estimated accurately by the IPCA method. In the following chapter, we elaborate on the state estimation procedure employed in IPCA and establish the link between this approach and the well-studied subject of Data Reconciliation.



(a)-1



(a)-2



(b)

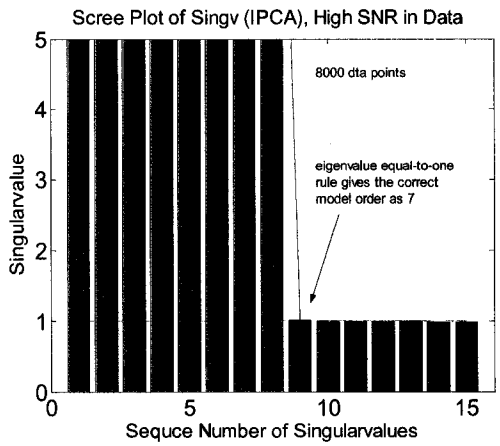
Figure 3.5 IPCA reveals correct model order

(a)-1: Cumulative variance plot in PCA, high SNR; model order is shown as 8~9.

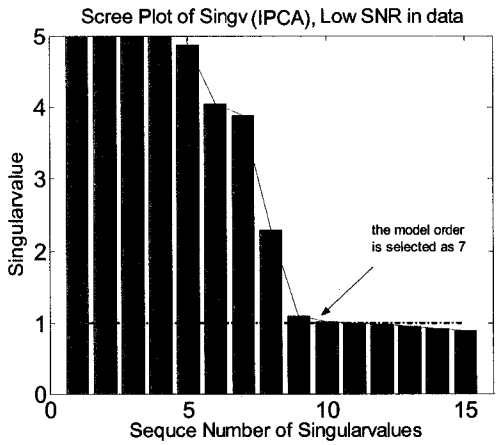
(a)-2: Scree plot in PCA, high SNR; model order is shown correctly as7.

(b): Scree plot in PCA, low SNR; model order is not clearly shown, but is shown as 8~9 according to eigenvalue-one rule.

(c) & (d): Scree plot in IPCA for high and low SNR data; model order is shown correctly as 7.



(c)



(d)

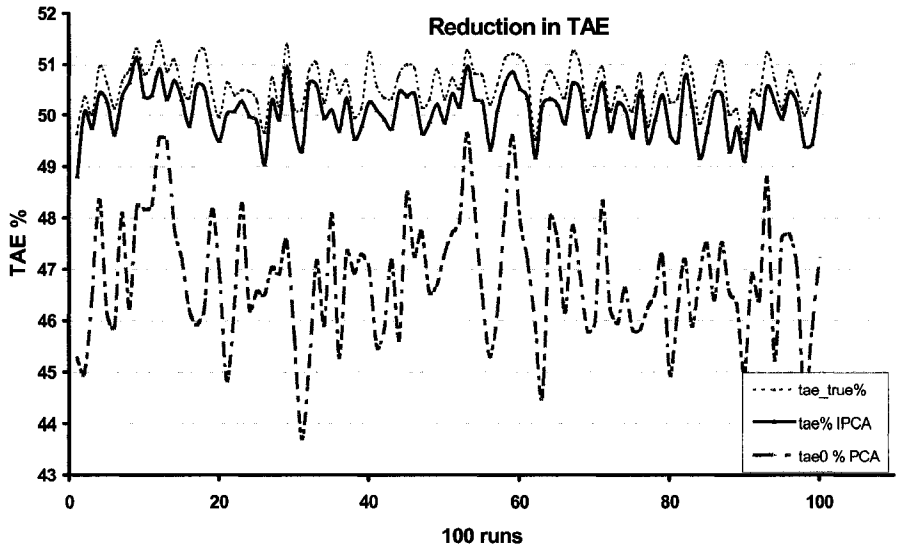


Figure 3.6. A comparison of the reduction of TAE obtained from the estimated models obtained by PCA and IPCA

Effect of Sample Size on Estimation Accuracy

Generally, it is well known that as sample size increases, the estimated model will be more accurate. In order to examine whether IPCA is able to provide consistent estimates, a series of simulations were carried out using data of different sample sizes but with the same set of signal and noise variations. The results of these simulations are presented in Table 3.2. It can be observed that as the sample size increase, more accurate estimates of the measurement error variances are obtained. except in the case of F15. This may be due to the fact that the measurement error variance for F15 is an order of magnitude smaller than the other error variances.

3.4 Conclusion

The IPCA approach is more appropriate for linear model identification than the PCA approach, especially when the errors in observation have much different variances. This is because PCA assumes that measurements of different variables have the same accuracy, which is not commonly the case. Monte Carlo analysis shows the advantage of IPCA approach in getting a good estimate of the process model, including a proper estimate of model order. The heuristic rules for model order determination in the PCA method gives ambiguous answers for model order in the example, even though they may do better by chance in some other cases. In addition, IPCA gives a good estimate of measurement errors. The model mismatch can be measured in two ways:

the angle between two subspaces; and the reduction in TAE that is obtained by data reconciliation using the estimated model. The reduction in TAE calculated from the true model is the limit that reveals the property of the system - the level of redundancy in the observed system. We cannot go beyond this limit.

F_i	F1	F2	F3	F4	F5	F6	F7
Std (F_i)	13.54	11.01	5.67	5.65	5.73	12.30	10.37
Std (e_i)	0.800	0.250	0.500	0.500	1.500	0.600	0.150
SNR	16.9	44.1	11.3	11.3	3.8	20.5	69.1
sample size	estimated standard deviations of measurement noise						
2000	0.777	0.268	0.496	0.512	1.457	0.577	0.154
6000	0.795	0.245	0.504	0.507	1.450	0.590	0.150
10000	0.805	0.245	0.501	0.509	1.474	0.599	0.150
14000	0.799	0.244	0.498	0.507	1.513	0.601	0.149
18000	0.805	0.236	0.498	0.505	1.521	0.602	0.149
22000	0.799	0.244	0.498	0.507	1.516	0.598	0.149
26000	0.800	0.243	0.500	0.509	1.518	0.599	0.150
30000	0.802	0.236	0.499	0.505	1.492	0.604	0.148
34000	0.799	0.245	0.496	0.504	1.476	0.602	0.147
38000	0.796	0.243	0.498	0.504	1.525	0.603	0.148
42000	0.800	0.249	0.497	0.503	1.522	0.598	0.148
46000	0.800	0.251	0.496	0.505	1.491	0.602	0.148

F_i	F8	F9	F10	F11	F12	F13	F14	F15
Std (F_i)	3.02	14.75	14.94	2.93	16.52	10.37	7.10	0.24
Std (e_i)	0.100	0.400	0.550	0.062	0.400	0.050	0.030	0.008
SNR	30.2	36.9	27.2	47.2	41.3	207.3	236.6	30.0
Sample size	estimated standard deviations of measurement noise							
2000	0.056	0.407	0.544	0.086	0.414	0.008	0.050	0.041
6000	0.072	0.404	0.549	0.066	0.408	0.038	0.008	0.027
10000	0.051	0.402	0.556	0.074	0.402	0.045	0.008	0.012
14000	0.061	0.403	0.557	0.068	0.401	0.047	0.025	0.015
18000	0.075	0.398	0.556	0.079	0.401	0.049	0.008	0.016
22000	0.081	0.402	0.555	0.069	0.401	0.049	0.008	0.011
26000	0.083	0.402	0.555	0.057	0.400	0.048	0.008	0.015
30000	0.086	0.399	0.556	0.068	0.401	0.048	0.008	0.022
34000	0.092	0.398	0.555	0.064	0.402	0.050	0.033	0.020
38000	0.087	0.397	0.554	0.069	0.403	0.047	0.037	0.025
42000	0.082	0.398	0.553	0.075	0.403	0.048	0.038	0.022
46000	0.079	0.399	0.552	0.070	0.403	0.049	0.043	0.021

Table 3.2 Larger sample size yields more accurate estimation of the noise variances

Chapter 4

Steady State Data Reconciliation (DR) Using Identified Model

4.1 Introduction

Efficient and safe plant operation can be achieved by monitoring key process variables that contribute to the economy of a process (e.g., yield of an operation) or are linked to equipment quality or status (fouling in a heat exchanger, activity of a catalyst), safety limits or environmental considerations. On the other hand, for on-line optimization, parameter estimation requires a set of reliable operational data. However, measurements of these variables are never error free and therefore the reconciliation of the data is necessary to estimate the state and condition of the plant.

In general, Data Reconciliation (DR), also called *validation*, aims at achieving the following two targets: (1) error reduction and (2) gross error (sensor biases and leaks) detection. The most commonly used formulation of DR is to minimize the sum of squares of the measurement corrections subject to model constraints and bounds. This technique allows us to adjust the measured data and to give estimates of unmeasured variables, where possible, in such a way that this set of reconciled data satisfies heat and material balance equations or *the constraint model.* This reconciled data can then be used for process monitoring, process analysis and evaluation, as well as process optimization and control.

The problem of DR was first brought up by Kuehn and Davidson in 1961. Vaclavek (1968; 1976) introduced ideas concerned with the treatment of unmeasured variables, and the optimal selection of measurements. Mah *et al.* (1976) derived the treatment of unmeasured variables and of gross errors. The concept of a projection matrix was proposed by Crowe *et al.* (1983) to eliminate unmeasured variables. The projection matrix can be obtained through QR factorization (Swartz, 1989; Sanchez and Romagnoli, 1996).

As for gross error detection and diagnosis, several test statistics and methods have been put forward. Methods based on steady-state linear data reconciliation include the *global test* (Reilly and Carpani, 1963), the measurement test (Mah and Tamhane, 1982; Crowe *et al.*, 1983), the nodal test (Reilly and Carpani, 1963; Mah *et al.*, 1976), the *generalized likelihood ratio* (GLR)

test (Narasimhan and Mah, 1987), the Bayesian test by Tamhane *et al.* (1988), *unbiased estimation* of *gross error* (UBET) (Rollins and Davis, 1992), and *principal component tests* (PCT) (Tong and Crowe, 1995). Gross error size estimation methods have been proposed by Madron (1985) and Narasimhan and Mah (1987). Multiple gross errors can be detected and eventually estimated through graph-theoretic analysis (Mah *et al.*, 1976) or serial elimination procedures or serial compensation procedures (Serth and Heenan, 1986). The *dynamic integral measurement test* was put forward by Bagajewicz *et al.* (1998) for dynamic DR problems, and GLR was again proposed by Narasimhan and Mah (1988). In this chapter, we will use IPCA to estimate the model and the error covariance matrix, and then apply *data reconciliation* to the flow network example.

4.2 Data Reconciliation (DR) - Problem Formulation

Given a measurement vector $\mathbf{y}$ of the true values of variables $\mathbf{x}$ and the measurement error covariance $\mathbf{\Sigma}_e$, with model $\mathbf{A}$ known where $\mathbf{Ax} = \mathbf{0}$, the problem of data reconciliation is one of reconstructing the observed data $\mathbf{y}$ by $\hat{\mathbf{x}}$ such that $\hat{\mathbf{x}}$ is an optimal estimate of the true value $\mathbf{x}$. The formulation, based on the MLE principle, can be written as shown below:

$$\min_{\hat{\mathbf{x}}} J_{DR} = (\mathbf{y} - \hat{\mathbf{x}})^T \mathbf{\Sigma}_e^{-1} (\mathbf{y} - \hat{\mathbf{x}}) \qquad \ldots\ldots\ldots (4.1)$$

$$st. \quad \mathbf{A}\hat{\mathbf{x}} = \mathbf{0}$$

The solution is:

$$\hat{\mathbf{x}} = \mathbf{y} - \mathbf{\Sigma}_e \mathbf{A}^T \left(\mathbf{A} \mathbf{\Sigma}_e \mathbf{A}^T \right)^{-1} \mathbf{A}\mathbf{y} \qquad \ldots\ldots\ldots (4.2)$$

Data reconciliation is based on measurement *redundancy* and conservation laws, making corrections on measurements. The reconciled values that come out of this procedure exhibit a lower variance compared to the original raw measurements; this allows the process to be operated closer to limits and therefore in a more efficient and economical manner.

Before doing DR, we need to know the process constraint model and the measurement error covariance. This formulation is based on the assumption that measurements have normally distributed random errors.

The presence of any gross errors (instrument biases and leaks) leads to incorrect estimates and severely biased reconciliation of the other measurements. To be able to perform DR efficiently, gross errors (or sensor faults), which behave differently than random noise, have to be filtered. Three central issues are of concern: detecting the presence of, identifying the location of, as well as estimating the size of, any gross errors. These topics are covered in chapter 5.

4.3 DR, PCA, and IPCA Filters

Equation 4.2 is a data filter that reduces the random error. PCA also filters out the noise in the residual space. As noted in chapter 2, we can make use of equation 4.3 to obtain filtered data by representing PCA as:

$$\hat{\mathbf{x}} = \sum_{i=1}^{k} v_i t_i = (\mathbf{I} - \hat{\mathbf{A}}^T \hat{\mathbf{A}})\mathbf{y} \qquad \ldots\ldots\ldots (4.3)$$

where $\hat{\mathbf{x}}$ is the filtered data.

The rows in $\hat{\mathbf{A}}$ are singular vectors corresponding to minor singular values. If the model $\hat{\mathbf{A}}$ is obtained from IPCA, the rows in $\hat{\mathbf{A}}$ are no longer orthogonal to one another and the filter cannot be written as equation 4.3. However, equation (4.3) can be applied to the scaled variables since PCA is applied to scaled data in the IPCA method. Thus we can write

$$\hat{\mathbf{x}}_s = \mathbf{y}_s - \hat{\mathbf{A}}_s^T \hat{\mathbf{A}}_s \mathbf{y}_s \qquad \ldots\ldots\ldots (4.4)$$

Substituting for the scaled variables, equation 4.4 can be written as

$$\hat{\mathbf{L}}^{-1}\hat{\mathbf{x}} = \hat{\mathbf{L}}^{-1}\mathbf{y} - \hat{\mathbf{L}}^T \hat{\mathbf{A}}^T \hat{\mathbf{A}} \hat{\mathbf{L}} \hat{\mathbf{L}}^{-1}\mathbf{y} \qquad \ldots\ldots\ldots (4.5)$$

where $\hat{\mathbf{L}}$ is the square root of the estimated covariance matrix $\hat{\mathbf{\Sigma}}_e$. Multiplying equation (4.5) by $\hat{\mathbf{L}}$ and noting that the rows of $\hat{\mathbf{A}}_s$ are orthonormal, equation 4.5 can be rewritten as

$$\hat{\mathbf{x}} = \mathbf{y} - \hat{\mathbf{\Sigma}}_e \hat{\mathbf{A}}^T \left(\hat{\mathbf{A}} \hat{\mathbf{\Sigma}}_e \hat{\mathbf{A}}^T \right)^{-1} \hat{\mathbf{A}}\mathbf{y} \qquad \ldots\ldots\ldots (4.6)$$

Equation 4.6 is identical to the reconciled estimates given by equation 4.2, with the estimated constraint model and estimated covariance matrix of errors being used. This shows that the IPCA method is closely allied with Data Reconciliation. In Data Reconciliation, the constraint matrix and measurement error covariance matrix are both assumed to be known. Typically, the constraint matrix is assumed to be derived from first principles, while the measurement error covariance matrix is assumed to be obtained from instruments manual or handbooks. Based on the above discussion, the IPCA method can be viewed as an approach for extracting this information from historical data.

Equation	Filter	Method	Condition
4.3	PCA	Total Least Square	Estimates **A** and assumes Σ_e to be an identity matrix
4.4	IPCA	Optimal Scaling	Simultaneously estimates **A** & Σ_e
4.2	DR	MLE	Assumes knowledge **A** & Σ_e

Table 4.1 A comparison of the three filters

4.4 Results and Discussion

In the preceding chapter, we compared the accuracy of the state estimates obtained using PCA, IPCA with the reconciled values (obtained under the assumption that both the constraint model and measurement error covariance matrix are known), using the measure of the *reduction in TAE* (Total Absolute Error). A bar chart of the reduction in TAE obtained by these methods for the 100 data segments used in simulation of the flow network is shown in Figure 4.1. The figure shows that data reconciliation using the IPCA model is effective. This is a consequence of the fact that the model and measurement error covariance matrix estimated by IPCA are good.

4.5 Concluding Remarks

To summary the above discussion we can arrive at the following conclusions:

- Data Reconciliation (DR) is a maximum likelihood filter.
- DR can be formulated as a real time problem.
- DR assumes that measurement error is multivariate normal.
- Requires that model **A** and the error covariance matrix Σ_e are given.
- If **A** and Σ_e are not known, they can be estimated by IPCA, and then used for DR.

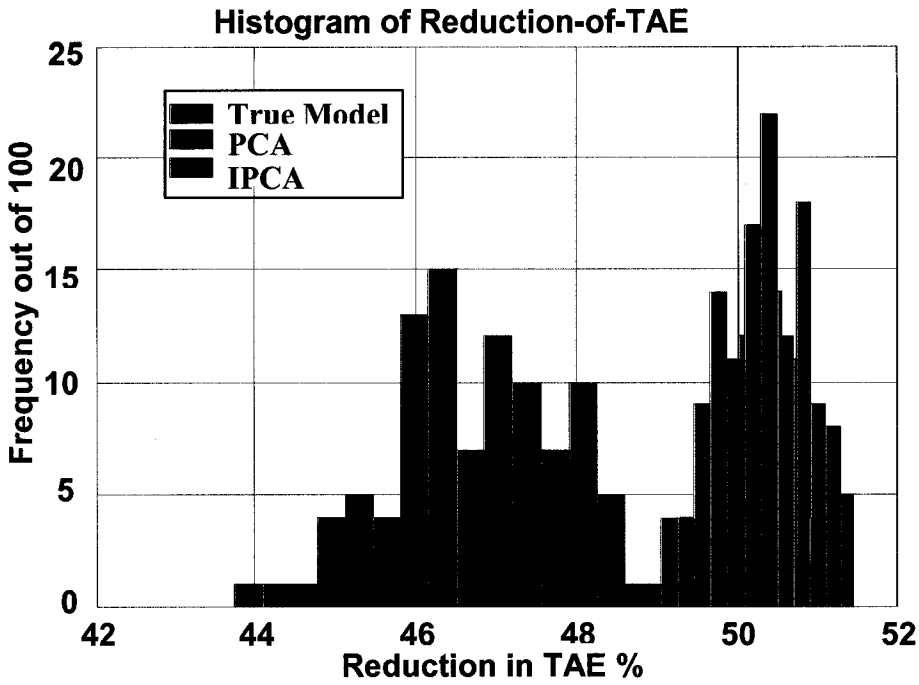


Figure 4.1 Applying the IPCA model in DR yields good results

Chapter 5

Sensor Fault Detection & Isolation Using Identified Model

5.1 Introduction

After a process model has been successfully identified, we can proceed with fault diagnosis. Earlier we have concluded that IPCA gives a better estimate of a process model. In this chapter we will focus on how to detect and isolate abrupt faults (typically sensor faults) using an identified model.

Faults can be classified into two types:

Type A fault: a fault that results in the violation of process constraints;

Type B fault: a abrupt and/or significant change in the process (often goes beyond the region of safe or economic operation) that will not violate process constraints (balances).

Type A faults can be further classified into two types (Yoon and MacGrrgor, 2001):

Simple fault: a fault that occurs in a specific fault source and only affects a single variable. This results when only one variable has changed its correlation pattern with all the remaining variables in the model. A *simple fault* is often caused by (1) malfunctioning or miscalibrated instrumentation (degradation of measurement accuracy, drifts, biases and so on), (2) sensor saturations or (3) process disturbances. In steady state, a simple *sensor fault* or an actuator bias results in a *simple fault.*

Complex fault: a fault associated with an abnormal change in the process such as (4) process leaks, (5) departure from steady state (if the process is monitored from a steady-state point of view), and (6) other process faults such as catalyst degradation, loss of reaction, valve stiction, equipment trips and even an urgent shutdown triggered by a safety protection control. In a process with feedback or feed-forward control loops, a *sensor fault* may result in a *complex fault* since the effects are propagated to other variables due to closed loops.

Type B faults usually results in high/low-limit alarms and has the potential to lead to a *type A fault.*

Any comprehensive fault diagnosis strategy should preferably possess the following capabilities (Narasimhan and Jordache, 2000):

- Ability to detect the presence of at least one fault in the data (the detection problem).
- Ability to identify the type and location of a fault (the identification problem).
- Ability to identify a fault that contains multiple *gross errors* simultaneously in the data (multiple gross error identification problem).
- Ability to estimate the magnitude of the sensor faults (the estimation problem).

In this chapter we will focus on the detection and identification of faults due to a single *gross error.*

After the constraint model is identified, it is used to monitor the abnormal events (faults) in the process that may lead the process to depart from its normal state. To be more explicit, we distinguish measurement errors or noise from faults. Process data is inherently inaccurate because of the underlying stochastic properties of the measurement errors. In fact any given process never conforms perfectly to its constraint model because of measurement noise and disturbance effects. Noise in observed data can cause small deviations of the observed process from its constraints. Note that noise does not only arise from stochastic sensor errors but also from minor disturbances and minor dynamic changes in the process. This is because we would never expect a process to be in a completely steady state. In model identification the training data is noisy and reveals "reasonable" process deviations from its constraints. On the other hand, the excitation of the process also defines a "normal" operation region, in which the operating point is expected to appear. As a consequence we can obtain statistical bases from the training data to monitor the process deviation both from its constraints and from its normal operating range. This is the starting point of all methods of fault detection.

Note that the identified constraint model itself may somehow present a mismatch not only due to poor data quality, but also as a result of the subtle process non-linearity and process change with time. Detailed discussion of this problem would lead us off in another research direction. To overcome the time-variant change in a process, some people suggest identifying the model and monitoring the operating process recursively (Li, *et al.*, 2000).

A number of statistical tests can be applied to detect the presence of any fault. Generally the outcome of hypothesis testing is not perfect. There are two types of testing errors:

Type I error: the statistical test detects the presence of faults, but in fact there is no fault.
Type II error: the statistical test fails to detect any existing fault in the data.

Type I error gives rise to a false alarm while *Type II error* means a missing alarm. The *power* of a statistical test is the probability of successfully detecting any existing fault such that:

$$power \text{ of a statistical test} = 1 - \Pr\{Type\ II\ error\}$$

where Pr means probability. For any statistical test, if the probability of *Type I error* is increased, the probability of *Type II error* will be decreased, and vice versa.

In this chapter we will first review the Q statistics of Squared Prediction Error (SPE) and T^2 statistics for fault detection, followed by the SPE contribution plot method for fault isolation. Later, *Global Test* (GT), *Squared Weighed Residual* (SWR) and *Generalized Likelihood Ratio* (GLR) tests are discussed. Then, three different schemes (including our proposed one) for fault detection and isolation are used in Monte Carlo simulations. The results show that our proposed method provides favorable performances for both detection and isolation of faults.

5.2 Fault Detection and Isolation (FDI)

5.2.1 Fault Detection Using T^2 and SPE

Two collective test statistics are widely used in Statistical Process Monitoring (SPM). First, Hoteling T^2 statistics indicates the variation within the process model in PCS. A large change in this subspace is observed if some points exceed the confidence limit in the T^2 chart, which means a big deviation (*type B fault*) occurs in the process. The other is the Q statistics, also known as the Squared Prediction Error (SPE), which monitors how well the data conforms to the constraint model. In the statistical sense, an unaccountable deviation from the normal model is observed if some points exceed the confidence limit in the SPE chart. In other words, this deviation means the breakdown of the correlation structure among the observed variables, which means a *type A fault*.

Researchers often introduce the PCA method followed by FDI methods using SPE and T^2 statistics, which seems to mean that **they go together**. However, PCA itself is just a method of identifying a non-causal correlation model of the process. After a model is obtained using PCA method, various methods for FDI can be applied. To perform SPE and T^2 analysis, we may use any model identified using various applicable methods other than PCA.

T^2 statistics

Traditionally, T^2 statistics is used as a collective test in various minimization processes, including data reconciliation. T^2 statistics can be calculated directly from the PCA representation.

$$\mathbf{T}_p = \mathbf{P}^T\mathbf{Y}$$

$$\left\{T^2\right\}_N = diag\left\{\mathbf{T}_p^T\left[\mathrm{cov}\left(\mathbf{T}_p\right)\right]^{-1}\mathbf{T}_p\right\} = diag\left\{\mathbf{Y}^T\mathbf{P}\left[\mathbf{P}^T\Sigma_y\mathbf{P}\right]^{-1}\mathbf{P}^T\mathbf{Y}\right\} \quad \ldots\ldots\ldots (5.1)$$

Where **P** ($n\times k$) is the matrix of principal loading vector columns, while $\left\{T^2\right\}_N$ is a $N\times 1$ vector.

This T^2 statistic measures the variation in PCS only. If we go on to assume that **Y** is multivariate normal, then, given that the actual mean and covariance from the population are known (or $N\rightarrow\infty$), the T^2 statistic threshold can be derived from

$$T_\alpha^2 = \chi_\alpha^2(k)$$

Here k represents the degree of freedom. If the actual covariance matrix is estimated from the sample matrix, the T^2 statistic can be derived as follows:

$$T_\alpha^2 = \frac{k(N-1)(N+1)}{N(N-k)}F_{\alpha,(k,N-k)}$$

We can also do the same thing in the residual space, in which case the T^2 statistic measures the deviation of data from the constraint model. We will discuss this point later.

Squared Prediction Error (SPE)

A plant-model mismatch, a *type A fault*, leads to significant prediction errors. The collective test, known as the *Q-statistics* or Rao-statistic, is designed to monitor the prediction error. Given the identified model $\mathbf{A} = \mathbf{B}^T$, we know that (refer to equation 2.5)

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{X}} = \mathbf{B}\mathbf{T}_b = \mathbf{B}\mathbf{B}^T\mathbf{Y} = \sum_{i=k+1}^{n} v_i v_i^T\mathbf{Y} \quad \ldots\ldots\ldots (5.2)$$

The SPE is then defined as:

$$SPE = diag(\hat{\mathbf{E}}^T\hat{\mathbf{E}}) \quad \ldots\ldots\ldots (5.2a)$$

SPE follows the Q-statistic. The distribution of the Q-statistic, as approximated by Jackson and Mudholkar (1979), is:

$$Q_\alpha = \theta_1\left[1 + \frac{h_o c_\alpha\sqrt{2\theta_2}}{\theta_1} + \frac{\theta_2 h_o(h_o-1)}{\theta_1^2}\right]^{1/h_o} \quad \ldots\ldots\ldots (5.3)$$

where $$h_o = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$$

$$\theta_l = \sum_{i=k+1}^{n} \lambda_i^l \qquad \text{for } l = 1, 2, 3.$$

and c_α is the confidence limit corresponding to the $(1-\alpha)$ percentile in a normal distribution. At a given level of significance, the threshold for the Q statistic can be calculated using the above equation and used to detect a *type A fault*. Although the derivation of this confidence limit expression **requires** that **Y** (=**X**+**E**) follow a normal distribution, it remains valid without the normality assumption as long as the error **E** is multivariate normal, because SPE is obtained from the residual space. That is, the "signal" power in the residual space is essentially acquired from the projection of the error **E**. From equation 5.2, we see that $\hat{\mathbf{E}}$ is obtained from projecting the residual score $\mathbf{T}_b = \mathbf{B}^T\mathbf{Y}$ in residual space back to the original data space.

Combination of T^2 and SPE

The T^2 and Q statistics detect different types of faults. Applying the two statistics together, we produce a cylindrical in-control region, as illustrated for $k = 2$ and $n = 3$ in the following Figure 5.1.

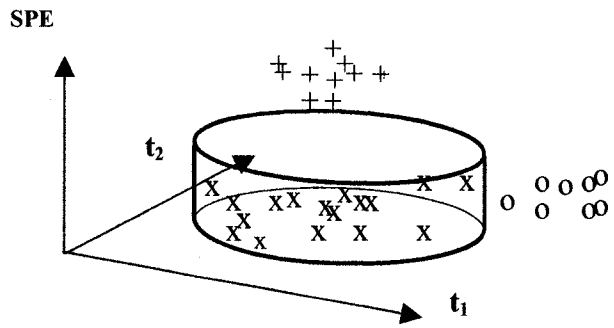


Figure 5.1 Visual expressions of SPE and scores

It is evident that the SPE and T^2 indices behave in a complementary manner. To simplify the fault detection task, Yue and Qin (2001) proposed a combined index φ for fault detection using SPE and T^2 as follows:

$$\varphi_j = \frac{SPE(\mathbf{y})}{Q} + \frac{T^2(\mathbf{y})}{\chi_k^2} = \mathbf{y}_j^T \mathbf{\Phi} \mathbf{y}_j, \quad j = 1,2, \ldots, N.$$

where $$\mathbf{\Phi} = \frac{\mathbf{P}\Lambda_p^{-1}\mathbf{P}^T}{\chi_k^2} + \frac{\mathbf{B}\mathbf{B}^T}{Q}$$

The distribution of φ can be approximated using $g\chi_h^2$, i.e.,

$$\varphi_j = \mathbf{y}_j^T \mathbf{\Phi} \mathbf{y}_j \sim g\chi_h^2$$

where the coefficient g is given by

$$g = \frac{tr\left(\Sigma_y \mathbf{\Phi}\right)^2}{tr\left(\Sigma_y \mathbf{\Phi}\right)}$$

and the degree of freedom for the χ^2 distribution is

$$h = \frac{\left[tr\left(\Sigma_y \mathbf{\Phi}\right)\right]^2}{tr\left(\Sigma_y \mathbf{\Phi}\right)^2}$$

5.2.2 Fault Detection Using GT and SWR

Global Test (GT)

Global test (GT) is a collective test that was first used in data reconciliation by Reilly and Carpani (1963). The optimal data reconciliation objective function values J_{DR} obtained from equation 4.1 are compared to $\chi^2_{1-\alpha,m}$, the critical value of distribution at the chosen level of significance α, where m is the rank of matrix **A**. If $J_{DR} \geq \chi^2_{1-\alpha,m}$, then H_0 is rejected and a gross error (fault) is detected. Note that $m < n$, where n = the number of variables. This is because the adjustments $\mathbf{a} = \mathbf{y} - \hat{\mathbf{x}}$ in data reconciliation are correlated and can be expressed in a reduced subspace, i.e., the *residual space* (RS). In fact, the data reconciliation adjustment is given from equation 4.2

$$\mathbf{a} = \mathbf{y} - \hat{\mathbf{x}} = \Sigma_e \mathbf{A}^T \left(\mathbf{A}\Sigma_e \mathbf{A}^T\right)^{-1} \mathbf{A}\mathbf{y}$$

where $\mathbf{a} \sim N(\mathbf{0}, \Sigma_e \mathbf{A}(\mathbf{A}\Sigma_e \mathbf{A}^T)^{-1}\mathbf{A}\Sigma_e)$, we find **a** is a linear transformation of a rank-reduced residual vector $\mathbf{r} = \mathbf{A}\mathbf{y}$.

The global test statistic is

$$\gamma = J_{\mathrm{DR}} = \mathbf{a}^T \Sigma_e^{-1} \mathbf{a} \sim \chi^2_{1-\alpha,\, m} \qquad \ldots\ldots\ldots (5.4)$$

It can be formulated in a more straightforward way (Ripps, 1965; Almasy, 1975; Madron, 1985):

$$\gamma = \mathbf{r}^T (\mathbf{A}\Sigma_e \mathbf{A}^T)^{-1} \mathbf{r} \sim \chi^2_{1-\alpha,\, m} \qquad \ldots\ldots\ldots (5.5)$$

Where $\mathbf{r} = \mathbf{A}\mathbf{y}$ is an $m\times 1$ residual vector, the corresponding observation is an $n\times 1$data vector $\mathbf{y}$. The γ value in equation 5.4 can be proved to be equal to that in equation 5.5.

Because of the great deal of research effort and intensive publications on the relative research topics, a brief review of them is given below: Corresponding to the collective test shown in equation 5.4, the univariate *measurement test* (MT) examines elements of **a**, which follows a standard normal distribution N (0,1); Corresponding to the collective test shown in equation 5.5, the univariate *constraint test* or *nodal test* (NT) examines elements of **r**, which also follows a standard normal distribution N (0,1). These two univariate tests will possess maximum power of detection (the probability of correct detection) when there is only one sensor fault at a specific time instant, if appropriate linear transforms are applied. Detailed derivations of GT were given by Crow *et al.* (1983, 1989), Mah and Tamhane (1982), and Almasy and Sztano (1975).

Compared to the univariate tests, the collective chi-square test detects not only sensor faults and leaks, but also system faults that result in any violation of process constraints. Tong and Crowe (1995) compared all the aforementioned statistical tests. They also defined a principal component transformation of the residual vector **r**, then performed a truncated chi-square test based upon the retained principal components. This method makes sense when the row rank of **A** is less than the number of rows in **A**.

The discussion given above assumes that model **A** and measurement error covariance Σ_e are already known. The following gives a method that performs GT with PCA modeling results.

Squared Weighted Residual (SWR)-A and Σ_e are unknown

Oxby and Shah (2000) pointed out the weakness of SPE for fault detection and isolation and went on to suggest a chi-square statistic, the Squared Weighted Residual (SWR) for the purpose of detection. From PCA or IPCA we can estimate model $\hat{\mathbf{A}}$ and the eigenvalue matrix Λ, enabling us to calculate the SWR value as

$$SWR = \mathbf{y}^T \hat{\mathbf{A}}^T \Lambda_b^{-1} \hat{\mathbf{A}} \mathbf{y} = \mathbf{r}^T \Lambda_b^{-1} \mathbf{r} \qquad \ldots\ldots\ldots (5.6)$$

Where SWR approximately follows a chi-square statistics with m degrees of freedom, where m is the selected model order. SWR in equation 5.6 is similar but not equivalent to γ in equation 5.5 due to process-model mismatch (which could be a mis-estimation of model order) or color noise. Even though SWR is only an approximation of chi-square, SWR-based fault detection is robust against the consistent model mismatch, because the weighting matrix Λ_b retains some model mismatch.

Note that the residuals **r** obtained from PCA are orthogonal and the residual space formed by **r** has full rank. If we apply PCA again in this residual subspace as proposed by Tong and Crowe (1995), the principal components will turn out to be the same as **r**. If PCA gives the exact estimation of model **A**, then SWR is very close to Tong & Crow's method. The only difference is that SWR does not truncate the residual space. Any truncation of a full rank residual space will result in the loss of redundancy in the data, and therefore is not recommended, because the redundant information in the data is valuable for fault detection and isolation. Fortunately the residual space is rarely singular in applications in industry.

Degrees of Freedom in Residual Space

Statistical fault detection and isolation relies on the redundancy in the data. The row rank of A measures this redundancy. In equation 5.6, the number of independent residuals that contribute equally to the value of the quadratic form (SWR) is the number of statistical degrees of freedom.

$$v_{swr} = rank(\Lambda_b) = m$$

According to the proceeding discussion, notice that $\mathbf{r} = \mathbf{Ay}$, we have

$$SPE_j = \hat{e}_j^T \hat{e}_j = r_j^T r_j$$

where $r_j \sim N(0, \lambda_b)$, we have

$$E\{SPE_j\} = \sum_{i=k+1}^{n} \lambda_{b,i} z_i^2 \qquad \ldots\ldots\ldots (5.7)$$

where $z_i \sim N(0,1)$. Hence SPE_j is a linear combination of chi-square variables of one degree of freedom at the jth time instant. In equation 5.7, some residuals will contribute very little if the eigenvalues are very different. This reduces the effective number of statistical degrees of freedom of SPE (Oxby and Shah, 2000; Box, 1954):

$$v_{spe} = \frac{tr(\Lambda_b)^2}{tr(\Lambda_b^2)} \le m \qquad \ldots\ldots\ldots (5.8)$$

Rewriting equation 5.8 we obtain:

$$v_{spe} = \frac{tr(\Lambda_b)}{\sum_{i=1}^{m}\left(\lambda_{b,i} \cdot \frac{\lambda_{b,i}}{tr(\Lambda_b)}\right)} = \frac{tr(\Lambda_b)}{\bar{\lambda}_b} \qquad \ldots\ldots\ldots (5.9)$$

From equation 5.9 we can see that the larger the eigenvalue, the more the corresponding residual vector contributes to the SPE value. SPE approximately follows a chi-square distribution with v_{spe} degrees of freedom, with the criterion estimated by equation 5.3. v_{spe} is a measure of the redundancy in the data that is effectively used by the SPE in fault detection and isolation. If all the small eigenvalues are identical, then SPE retains all the useful information and is equivalent to SWR because $v_{spe} = v_{swr} = m$. If we consider another extreme case in which the largest eigenvalue in Λ_b is much larger than others, we can see that v_{spe} =1 and most of the useful information is ignored. The small eigenvalues typically decrease geometrically so that the value of v_{spe} is commonly 1/3~1/2*m*. This means that one-half to two-thirds of the diagnostic information in the data is commonly lost in using SPE.

5.2.3 Fault Isolation

Introduction

Once the fault has been detected, the next step is expected to determine its cause and location. However, isolating the fault is a challenging task for plant operators and engineers because a large number of process variables are usually monitored. The objective of fault isolation is to determine which plant variables have contributed to the observed out-of-control behavior, thereby letting the operators and engineers focus on the subsystem where the fault occurred. In this way, fault isolation can effectively help the operators and engineers in the process-monitoring scheme and therefore significantly reduce the risk of safety problems or losses in profitability.

A number of approaches to fault isolation are employed within MSPC, including examining contributions to the T^2 statistic (Wise and Gallagher, 1996), contributions to SPE (Miller *et al.*, 1993), contributions to individual scores (Montague *et al.*, 1998), model prediction errors (Mah and Tamhane, 1982; Dunia *et al.*, 1996), Maximum Power (MP) modification on Measurement Test (MT) and Nodal Test (NT) (Crow *et al.*, 1989; Mah and Tamhane, 1982, 1985; Almasy and Sztano, 1975; Rollins *et al.*, 1996), and partial correlations or similar approaches (Ibrahim and Tham, 1995). A residual space decoupling method, the Principal Component Test (PCT), has been proposed by Tong and Crowe (1995, 1996) as an alternative to multiple measurement bias and leak identification. Tong and Bluck (1998) reported industrial applications of this method to illustrate that PCT is more sensitive to subtle gross errors than are other methods and has greater power to correctly isolate the sensor faults than the conventional Nodal, Measurement and Global

Test (NT, MT and GT). However, PCT has proven to be less efficient than MT when used for multiple sensor fault identification (Bagajewicz *et al.*, 1999, 2000). To isolate faults, Narasimhan and Mah (1987) introduced the Generalized Likelihood Ratio (GLR) test, which can also provide an estimate of the size of the isolated sensor fault. They showed, as did Crowe (1988), the equivalency between GLR and the Maximum Power Measurement Test (MP-MT) when there is only one sensor in fault. Yue and Qin (2001) introduced a reconstruction-based method to identify unidimensional and multidimensional faults. More complex classification approaches can be used if we predefine and carefully study the fault conditions (Hand, 1981, 1982). The methods cited above are basically for linear steady state process monitoring and fault diagnosis. Literature surveys can be found in Narasimhan and Jordache (2000) and in Sánchez and Romagnoli (2000). The discussion on model-based analytical redundancy approach to FDI, which is applied to dynamic non-linear systems, is beyond the scope of this paper. Surveys on this topic can be found in Frank (1990), Gertler (1988), and Chen and Patton (1999).

Next we will discuss the most commonly used approaches for fault isolation: the contribution methods and GLR method.

Score Contribution Method

The Score Contribution approach is to evaluate the contribution of each process variable y_i to an individual score $t_j = p_j^T \mathbf{y}$, then sum up these contributions to only those scores that violate their individual thresholds. The univariate *constraint test* or *nodal test* (NT) examines values of t_j for the purpose of detection. The threshold is determined under a certain level of overall *type I error.* A conservative estimate of the probability α (typically $\alpha = 0.05 \sim 0.01$) of overall *type I error* is given by Sidak (1967) and, accordingly, the probability of any individual *type I error* can be obtained as:

$$\beta = 1 - (1-\alpha)^{1/k} \qquad \ldots\ldots\ldots (5.10)$$

where k is the dimension of the principal component space and also the number of scores t_j ($j = 1 \sim k$). Then the threshold for each of the normalized scores is $Z_{1-\beta/2}$, where scores are normalized by their standard deviations. This is a conservative threshold because scores t_j may correlate with one another, unless they are orthogonal, as are those obtained from PCA.

Better performances for detection are attained if we form a *maximum power test* (MP test) (Crowe, 1989):

$$t_j^* = \frac{\left|\left(\mathbf{W}^{-1}\mathbf{P}^T\mathbf{y}\right)_j\right|}{\sqrt{\left(\mathbf{W}^{-1}\right)_{jj}}} \sim N(0, \beta) \qquad \ldots\ldots\ldots (5.11)$$

where $\mathbf{W} = \mathrm{cov}\left(\mathbf{P}^T y\right) = \mathbf{P}\Sigma_e\mathbf{P}^T$.

This test is valid under the assumption that scores are normally distributed.

The sensitivity for individual scores is defined as:

$$s_{t_j, y_i} = \frac{\partial(P_j^T\mathbf{y})}{\partial y_i} = p_{ij} \qquad \ldots\ldots\ldots (5.12)$$

From this equation, we can see that the sensitivity to a given variable y_i is the loading of this variable for a given score t_j. Instead of looking only at the loadings (sensitivities), the score contribution method looks at both the variables and the loadings. The difference between contributions and loadings becomes significant when some of the process variables have a value close to zero. In such a case, those same variables may have large loadings, but the contributions are very small. The procedure of the score contribution method is applied as follows:

(1) Check the normalized scores t_j^2 / W_{jj} for the observation **y** and determine the $l \le n$ scores responsible for alarms in detection procedures such as GT or MP nodal test.

(2) Calculate the contribution of each variable y_i to the out–of-control scores t_j.

$$cont_{t_j, y_i} = \frac{t_j}{W_{jj}} p_{ij}\left(y_i - \mu_i\right)$$

where p_{ij} is the (i,j)th element of the loading matrix **P**.

(3) When $cont_{t_j, y_i}$ is negative, set it equal to zero.

(4) Calculate the total contribution of the variable, y_i.

$$CONT_{yi} = \sum_{j=1}^{l}\left(cont_{t_j, y_i}\right) \qquad \ldots\ldots\ldots (5.13)$$

(5) Plot $CONT_{yi}$ for i =1~n process variables and over a certain sample length of data.

The variables responsible for the fault can be prioritized or ordered by the total contribution values, enabling the plant operators and engineers to immediately focus on those variables.

The individual score contribution approach is essentially a heuristic method. This is because the contributions may have different signs, there will be cancellation effects among the contributions from different variables. Despite this disadvantage, people can study the contributions by checking both the signs and magnitudes to obtain hints of the cause of a fault alarm when the process is not that complicated. Tong and Crowe (1995) successfully identified a process leak using Score Contribution Plots in which the scores were obtained in the mass-balance residual space.

T^2 Contribution Method

Contributions to the T^2 statistic are obtained by taking the gradient of T^2 with respect to each variable. From equation 5.1 the gradient of the T^2 statistic gives the sensitivity to the variable vector **y** as:

$$\mathbf{s}_{T^2} = \frac{\partial(T^2)}{\partial \mathbf{y}} = 2\mathbf{P}\left(\mathbf{P}^T\Sigma_y\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{y} = 2\mathbf{P}\Lambda_p^{-1}\mathbf{P}^T\mathbf{y} \qquad \ldots\ldots\ldots (5.14)$$

we can ignore the constant "2" and then the contribution from each variable is

$$T^2CONT_{y_i} = \int_0^{y_i}(\mathbf{s}_{T^2})_i dy_i = \sum_{j=1}^{k}\left(p_{ij}y_i\right)^2/\lambda_j = y_i^2\sum_{j=1}^{k}\left(p_{ij}^2/\lambda_j\right) \qquad \ldots\ldots\ldots (5.15)$$

If $T^2CONT_{y_i}$ is the largest in all values calculated for $i = 1 \sim n$, then y_i is indicated as a potential cause of the fault.

SPE Contribution Method

After an alarm shows up in the SPE chart, the SPE contribution plot can be used to isolate the fault. Let y_j ($n\times 1$) denote the observations whose true value can be predicted from the PCA model as $\hat{x}_j = \mathbf{P}\mathbf{P}^T y_j$, where **P** is the matrix of retained principal component vectors. The error vector at time instant j is given by $\hat{e}_j = y_j - \hat{x}_j = \mathbf{B}\mathbf{B}^T y_j$, and from equation 5.2a, we have:

$$SPE_j = \hat{e}_j^T\hat{e}_j$$

The fractional contribution of the ith variables to the overall *SPE* at sample instant j can be computed as (Miller *et al.*, 1993):

$$SPECONT_{y_i} = \frac{\hat{e}_{ij}^2}{SPE_j}, \ (i = 1, 2, \ldots, n; j = 1, 2, \ldots, N) \qquad \ldots\ldots (5.16)$$

By plotting a group of these values, we can identify the variables making a significant contribution to SPE as the suspected cause of the fault. From previous discussions, it is already known that useful diagnostic information (the degree of data redundancy) will be lost in implementing SPE. This is why we sometimes observe misleading results obtained from SPE contribution plots. Yue and Qin (2001) have given an example in which this method is invalid.

Generalized Likelihood Ratio (GLR) Method

Willsky and Jones (1974) developed the GLR method to identify abrupt failures in dynamic systems. Narasimhan and Mah (1987) introduced the General Likelihood Ratio (GLR) test for isolating gross errors, a test that can also estimate the magnitude of the isolated gross errors. They also proved that the GLR test and the maximum power MT are equivalent for identifying sensor faults. Bagajewicz and Rollins (2003) provided proof that when there is only one sensor bias, the maximum power MT and GLR tests are consistent from a statistical point of view, which means one can identify a fault correctly under deterministic conditions.

Given that the balance model **A** is already known from the knowledge of the process (such as the process shown in Figure 3.1), the balance residuals are

$$\mathbf{r} = \mathbf{A}\mathbf{y} = \mathbf{A}(\mathbf{x} + \mathbf{e}) = \mathbf{A}\mathbf{e}$$

If no gross errors are present, we have

$$E\{\mathbf{r}\} = \mathbf{0}$$

$$\mathrm{cov}\{\mathbf{r}\} = \mathbf{A}\Sigma_e\mathbf{A}^T = \mathbf{W}$$

If a sensor bias of magnitude b is present in measurement i, then

$$\mathbf{y} = \mathbf{x} + \mathbf{e} + b\mathbf{e}_i \qquad \ldots\ldots\ldots (5.17)$$

$$E\{\mathbf{r}\} = b\mathbf{A}\mathbf{e}_i \qquad \ldots\ldots\ldots (5.18)$$

where $\mathbf{e}_i$ is a vector with unity in position i and zero elsewhere.

If a process leak of magnitude b is present in unit (node) j, then

$$\mathbf{A}\mathbf{x} - b\mathbf{m}_j = \mathbf{0} \qquad \ldots\ldots (5.19)$$

$$E\{\mathbf{r}\} = b\mathbf{m}_j \qquad \ldots\ldots (5.20)$$

where the jth element of unit vector $\mathbf{m}_j$ is in unity. $\mathbf{m}_j$ may take different values if we are looking at energy balance or component mass balance (Narasimhan and Mah, 1987). Equation 5.18 and 5.20 may be simplified as

$$E\{\mathbf{r}\} = b\mathbf{f}_k \qquad \ldots\ldots (5.21)$$

$$\text{where } \mathbf{f}_k = \begin{cases} \mathbf{A}\mathbf{e}_i & \text{for a sensor bias in measurement } y_i \\ \mathbf{m}_j & \text{for a process leak in unit } j \end{cases} \qquad \ldots\ldots (5.22)$$

We call $\mathbf{f}_k$ the *fault signature vector.*

The derivation of GLR begins with the hypotheses for gross error detection:

$$\begin{aligned} H_o &: E\{\mathbf{r}\} = \mathbf{0} \\ H_1 &: E\{\mathbf{r}\} = b\mathbf{f}_k \end{aligned} \qquad \ldots\ldots (5.23)$$

The likelihood ratio test statistic is given as

$$\gamma = \sup \frac{\Pr\{\mathbf{r}|H_1\}}{\Pr\{\mathbf{r}|H_0\}} = \sup_{b,\mathbf{f}_k} \frac{\exp\left[-\frac{1}{2}(\mathbf{r} - b\mathbf{f}_k)^T \mathbf{W}^{-1}(\mathbf{r} - b\mathbf{f}_k)\right]}{\exp\left[-\frac{1}{2}\mathbf{r}^T\mathbf{W}^{-1}\mathbf{r}\right]} \qquad \ldots\ldots (5.24)$$

Rewriting equation 5.24 we have

$$T = 2\ln(\gamma) = \sup_{b,\mathbf{f}_k}\left[\mathbf{r}^T\mathbf{W}^{-1}\mathbf{r} - (\mathbf{r} - b\mathbf{f}_k)^T \mathbf{W}^{-1}(\mathbf{r} - b\mathbf{f}_k)\right] \qquad \ldots\ldots (5.25)$$

In equation 5.25, for every choice of fault signature vector $\mathbf{f}_k$, we can obtain the maximum likelihood estimate of b as

$$\hat{b} = \left(\mathbf{f}_k^T\mathbf{W}^{-1}\mathbf{f}_k\right)^{-1}\left(\mathbf{f}_k^T\mathbf{W}^{-1}\mathbf{r}\right) \qquad \ldots\ldots (5.26)$$

Substituting $\hat{b}$ into equation 5.25 we get

$$T_k = \frac{\left(\mathbf{f}_k^T \mathbf{W}^{-1} \mathbf{r}\right)^2}{\mathbf{f}_k^T \mathbf{W}^{-1} \mathbf{f}_k} = \frac{d_k^2}{C_k} \qquad \ldots\ldots (5.27)$$

and

$$T = \sup_k T_k \qquad \ldots\ldots (5.28)$$

The fault k is then isolated if T_k gives the largest value in the GLR test.

In equation 5.27 notice that $d_k \sim N(0, D_k)$ under H_0, and consequently T_k has a central χ^2 distribution with one degree of freedom. In regard to equation 5.10 we may choose the critical value $\chi^2_{1-\beta,1}$ as the threshold for the test. If T overshoots this critical value, a gross error is detected. In applying this procedure, we cannot avoid looking at two issues:

First, as mentioned before, the constraint model **A** may be easily obtained as balance equations from the process schematic so that we know how a process leakage affects the balance. However, if we look at a complex process, where model **A** is estimated as $\hat{\mathbf{A}}$ from PCA, IPCA or other identification methods, then we should find the *fault signature vector* for leakage instead of simply assigning it as m_j in equation 5.22. This is because leakage of material at one unit j will contaminate all constraints. Equation 5.19, 5.20 are modified as

$$\hat{\mathbf{A}}\mathbf{y} - b\hat{\mathbf{A}}\mathbf{G}_j = \hat{\mathbf{A}}\mathbf{e}$$

and

$$E\{\mathbf{r}\} = b\hat{\mathbf{A}}\mathbf{G}_j \qquad \ldots\ldots (5.29)$$

where $\mathbf{G}_j$ is the balance-equation-coefficient vector at unit j and is normalized to unity. Accordingly the fault signature vectors are obtained as

$$\mathbf{f}_k = \begin{cases} \hat{\mathbf{A}}\mathbf{e}_i & \text{for a sensor bias in measurement } y_i \\ \hat{\mathbf{A}}\mathbf{G}_j & \text{for a process leak in unit } j \end{cases} \qquad \ldots\ldots (5.30)$$

Second, in DR and GED (gross error detection), we consider only sensor bias and process leaks or loss of energy somewhere in the process. However, the fault may be a kind of abnormal process status that is not caused by sensor bias and leaks. In this case we have two options:

(1) By applying a GED strategy directly, such as GT, MP-MT or GLR test, the fault isolation results will point out all process variables suspected of containing a "sensor fault." We

can then analyze the isolation report and use our process knowledge to make an engineering judgment to determine the most likely problem.

(2) Build a *fault bank* that contains all the pre-specified possible faults and then derive a group of fault signature vectors for the GLR method. If we cannot do this analytically, we can numerically estimate the fault signature vectors through training.

From the above discussion, we understand GLR as summarized below:

- The GLR test can easily detect the existence of gross errors and isolate the source of those errors.
- The consistency of GLR for fault isolation can be proved for the case of a single sensor fault.
- Assuming the process constraint model **A** is perfect and measurement noises are multivariate normal, the individual GLR test value for each fault signature vector follows a central chi-square distribution with one degree of freedom.
- Before performing the GLR test, we need to model all possible faults that are prespecified based upon an understanding of the process. If a fault cannot be modeled either analytically or numerically, it cannot be isolated by GLR in a quantitative way.

5.2.4 Adjustability and Detectability

Adjustability

All methods in DR and FDI rely on the redundancy of the information obtained from the observed process. The information redundancy can be measured by adjustability (Madron, 1992).

$$a_i = \left(1 - \frac{\sigma_{\hat{e}_i}}{\sigma_{e_i}}\right) > a_{\mathrm{CR}} \qquad \ldots\ldots (5.31)$$

where a_i is the adjustability of the ith variable, $\sigma_{\hat{e}_i}$ is the standard deviation of $\hat{e}_i$ that is obtained from DR or IPCA filtering, and σ_{e_i} is the standard deviation of measurement error. a_{CR} is a selected critical value from interval (0,1). If a_i <0.1, for instance, data reconciliation or IPCA filtering will not significantly improve the accuracy of measurement i, and the adjustment made to this variable will also be small.

Detectability

Charpentier *et al.* (1991) make note of a factor that measures the *detectability* of an error.

$$d_i = \sqrt{1 - \frac{\sigma_{\hat{e}_i}^2}{\sigma_{e_i}^2}} \qquad \ldots\ldots (5.32)$$

This factor also measures the redundancy of the measurements. The more redundant the measurement i is, the greater the contribution of the error in measurement i to the constraint imbalances. Therefore, the gross error is more likely to be detected, which means a larger detectability factor. This implies that if the same value of the statistical test (e.g., GLR) is obtained from more than one measurement, the ones that have large detectability factors are likely to be contaminated by errors of small magnitudes. Some measurements commonly approach non-redundancy, in which case their detectability factors, relatively speaking, are very low. Jordache (1985) and Charpentier *et al.* (1991) have carried out simulation studies on the redundancy problem. In the flow-network in Figure 3.1, for example, if we obtain an estimate of model **A** and the measurement error covariance, we can estimate the detectability of all the measurements so that we will know in advance which sensor fault is likely to be detected and which is not. Table 5.4 shows the estimation results for the flow-network example.

In a simulation example, negative detectability may be obtained due either to a model estimation mismatch or an over-estimation of the true measurement noise. This over-estimation is possible when the process disturbance (often color noise) is assigned to measurement noise.

5.3 Sensor Fault Detection and Isolation

If the process can be well described by a local linear model, then this model can be obtained from PCA or IPCA modeling. We can apply FDI schemes based on the identified model and the estimated error covariance matrix. This section will compare FDI performances of different combinations of model ID and FDI schemes via Monte Carlo simulations.

5.3.1 Fault Diagnosis Strategies

PCA-SPE

This method uses PCA to identify the process constraint model, with model order determined by the eigenvalue-one rule. It also uses Q-statistics for detection and SPE contribution plot for isolation.

PCA-SWR-GLR

This method uses PCA to identify the process constraint model, with model order determined by eigenvalue-one rule. It also uses SWR statistics for detection and GLR statistics for isolation.

IPCA-SWR-GLR

This method uses IPCA to identify the process constraint model and measurement covariance matrix, with the model order determined by eigenvalue equal-to-one rule. It furthermore uses SWR statistics for detection and GLR statistics for isolation. In this method, the detectability of

all variables can also be estimated, which reveals knowledge of which variables will easily be detected as having a fault.

5.3.2 False Alarm Rate and Threshold Adjustment

Before proceeding with our comparison of the FDI performances of different strategies, we need a basis that will make it possible to carry out a fair comparison of the methods concerned. This issue is associated with the determination of an appropriate threshold for fault detection. Any increase in the threshold certainly reduces the *type I error* for detection tests and at the meantime reduces the sensitivity of the tests. All the tests would be expected to have the same sensitivity for the fault-free data. If the statistical thresholds for the different tests were ideally accurate, this condition would be automatically satisfied, but in practice, this is not the case. Recall that the confidence limits obtained from Q statistics require the assumption that the observed values are temporally non-correlated and normally distributed. In practice, this assumption hardly ever holds. There are various reasons leading to the degradation of the statistical base for threshold determination for Q test, SWR test, and Hotelling's T^2 test:

(1) Process disturbances or colored noise leads to auto-correlation of the residual. However, the deviation of residuals from the normality assumption can be ignored to some extent due to the central limit theorem, and the statistical thresholds are commonly acceptable (Wise *et al.*, 1990);
(2) Highly dynamic signals break down the assumption of independence, resulting in the distortion of the T^2 threshold;
(3) Over estimation of model order makes the model recognize part of signals (which are often auto-correlated) as noise. This means that extra power within specific bandwidths will be added into the residual space, leading to a colored residual.
(4) Any process deviation form linearity assumptions would certainly affect the thresholds;
(5) In practice, the training data is not totally free of any minor faults or abnormal episodes; For steady state study, the training data is hard to be absolutely stationary.
(6) Given the limit of sample size, the estimated model cannot perfectly fit the observed process.

Special care should be taken when looking at the T^2 chart in principal component subspace (PCS), because the process signals are commonly auto-correlated. However, the statistical tests in the residual space (RS) can be easily formulated under the Gaussian assumption as long as the noises are white and the residual space is decoupled sufficiently from the true signal space. Conventional PCA, as we already known from chapter 2, does not give an accurate model order

and therefore can easily lead to threshold degradation for FDI schemes performed in the residual space.

Before long we will perform Monte Carlo simulations in which we will train the threshold for detection purposes. We can define the *false alarm rate* for training data as:

$$\upsilon = \frac{\text{number of false alarm incidents}}{\text{number of fault free observations}} \qquad \ldots\ldots (5.33)$$

Improper choice of thresholds leads to significant inflation and shrinkage of the *false alarm rate*. If the desired *false alarm rate* is set at 1% for the training data, we can adjust the thresholds by the trial and error method until all test runs (100 runs in the Monte Carlo simulations) using different model ID and FDI methods have the same value of π. Having established a fair basis of comparison, we can then compare the *fault detection rates* and *correct isolation rates* simulated from different methods.

In the simulations we define:

$$\text{fault detection rate} = \frac{\text{number of samples that are detected as faulty}}{\text{number of fault contaminated samples (1000)}}$$

$$\text{correct isolation rate} = \frac{\text{number of samples where faults are correctly isolated}}{\text{number of samples that are correctly detected as faulty}}$$

5.3.3 Simulation Results and Discussion

To illustrate and compare the performances of three different FDI schemes presented in Section 5.3.1, we take the flow-network example and perform Mont-Carlo simulations. This example limits the discussion to the application of these schemes for the sensor fault detection problem. Even in this simplified condition, there are still some open problems and ample room for improvement.

Figure 5.2 represent the key procedures of the simulation.

Generate Fault-Contaminated Data

To generate the fault-contaminated data, we randomly select one of the 100 segments of the fault-free data, for example, the first segment of the data, and then introduce a sensor bias (may vary for different FDI examples) to the 1001~2000th observations of the data. Figure 5.3 shows how the fault-contaminated data is generated, and for the *i*th observation we have

$$y_i = x_i + e_i + b_i$$

where

- e_i is the white measurement error, i.e., $e_i \sim N(0, \Sigma_e)$, where Σ_e is diagonal
- x_i is the true value of the observations (slow, colored signals)
- b_i is the vector of sensor bias added to the data, here for i =1001~2000, b_i remains the same.

Perform the Fault Detection and Isolation Using Different Schemes

The following steps were used for this example:

(1) Generate training data and simulate different sensor faults as described above. Two sets of data with high SNR and low SNR are generated for the purpose of comparison of PCA and IPCA modeling performances. We select high SNR data, as the simulation data to proceed with further analysis.

(2) Perform PCA modeling, using the auto-scaling method to scale the data before the analysis.

(3) Perform IPCA modeling and compare the results via the data reconciliation efficiencies with those obtained from PCA. This work has been shown in Section 3.2.

(4) Check the detectabilities of different measurements according to the system properties we estimated in step (3). Choose measurement F4 (with high detectability) and F11 (with very low detectability) as fault contaminated variables.

(5) Adjust thresholds in SPE, SWR and T^2 charts using training data to obtain a fair basis for the comparison as discussed in Section 5.3.2.

(6) Compare sensor fault detection and isolation results using the three fault diagnosis strategies presented in Section 5.3.1. In doing this, the following scenarios are considered:

- When both PCA model order and IPCA model order are correctly selected.
  - Different sensor faults (in F4 and F11);
  - Different sensor fault sizes (in F11).
- When PCA model order is different from IPCA model order.
- Different model orders, but set $\mathbf{W} = \mathbf{A}\Sigma_e\mathbf{A}^T = \mathbf{I}$ in PCA/GLR approach for isolation.

(7) Check the effect of sample size on error covariance estimation in IPCA modeling.

Two sets of training data were generated with high and low SNR respectively (refer to Table 6.1) and then PCA and IPCA were applied to them, yielding the model identification results as illustrated in Figure 3.5. For data with high SNR values, PCA provides an adequate determination of the correct model order (Figure 3.5 (a)-1, (a)-2), but becomes inefficient for data with low SNR

values (Figure 3.5 (b)). IPCA gave good results no matter whether the SNR values were high or low (Figure 3.5 (c) & (d)).

The eigenvalue equal-to-one rule was applied in IPCA, and we can find in Table 5.1 that the model order is 7. The estimated standard deviations of error (noise) obtained from the IPCA method are listed in Table 5.2. Note that the estimations are close to the true values. The more redundancy the process has, the more accurate we expect the estimation to be.

Recall that in Chapter 3, Figure 3.3 has compared model identification results obtained from PCA and IPCA. Model estimation errors (angles between $\mathbf{A}$ and $\hat{\mathbf{A}}$) are smaller and more consistent using IPCA than using PCA. Similarly, Figure 3.6 has shown that a greater amount of TAE reduction is obtained from IPCA than from PCA.

Table 5.3 gives the adjustability and detectability obtained from the true mechanistic model of the process and true measurement noise generated in simulation. We notice that Table 5.4 gives similar results that are obtained from the process information that is estimated using IPCA. As mentioned before, we choose measurement F4 (with high value of detectability index = 0.978) and F11 (with very low value of detectability index = 0.176) as fault contaminated variables.

Figures 5.4 and 5.5 compare the detection rate and isolation rate using the PCA-SPE and IPCA-SWR-GLR methods. Constant bias type faults of size $+4\sigma$ are added to flow F4 that has a high detectability index. The model order selected in PCA is supposed to be the same as in IPCA, with the correct number being 7.

A group of sensor faults of different sizes are also added into flow F11. The FDI results in this case are shown in Figure 5.6. When the size of fault $<10\sigma$, neither PCA-SPE method nor IPCA-SWR-GLR method give high detection rate and isolation rate. This is because the detectability of flow F11 is only 0.176, much lower than that of F4 (0.978). When the size of the sensor fault becomes greater than 20σ, we observe that IPCA-SWR-GLR method gives reasonably good results for FDI. Notice that in flow F11, $20\sigma = 1.33$ (refer to Table 5.2) still represent a relatively small change to the process, if compared to a change of size 2σ in flow F4 that is 1.01.

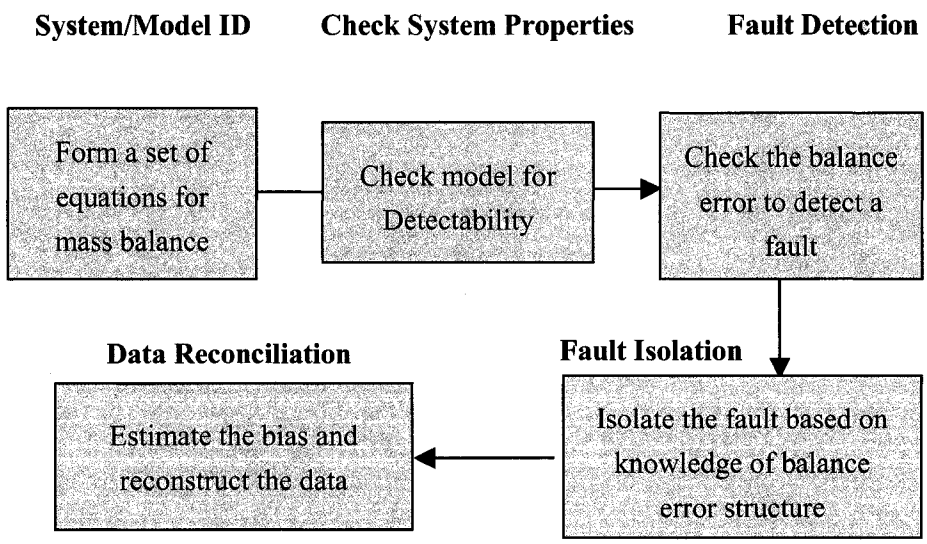


Figure 5.2 The procedure adopted in this thesis for fault detection and isolation

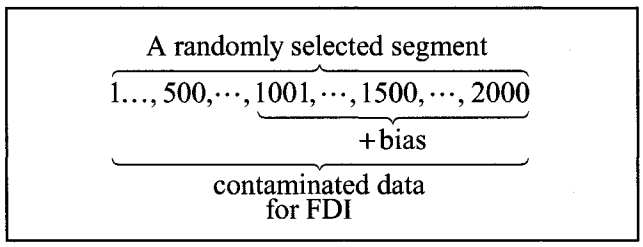


Figure 5.3 One segment of sensor fault-contaminated data for FDI

No.	Eignv_IPCA
1	116144268
2	54955
3	42138
4	4861
5	1930
6	954
7	428
8	174
9	1.03
10	1.01
11	1.00
12	1.00
13	0.99
14	0.98
15	0.97

Table 5.1 Eigenvalues obtained from IPCA, which verify the model order being 7 (High SNR data set)

TAG	std_n_ true	std_n_est
F1	0.8000	0.8035
F2	0.2500	0.2404
F3	0.5000	0.5036
F4	0.5000	0.5045
F5	1.5000	1.4585
F6	0.6000	0.5942
F7	0.1500	0.1500
F8	0.1000	0.0812
F9	0.4000	0.3991
F10	0.5500	0.5587
F11	0.0616	0.0663
F12	0.4000	0.4069
F13	0.0500	0.0458
F14	0.0300	0.0281
F15	0.0080	0.0081

Table 5.2 Error covariance estimated by IPCA for high SNR data
(std_n_true ---- simulated standard deviation of noise;
std_n_est ---- estimated standard deviation of noise)

TAG	**Error Variance**	Adjustability	Detectability
F1	0.6400	0.551	0.894
F2	0.0625	0.088	0.410
F3	0.2500	0.357	0.766
F4	0.2500	0.766	**0.972**
F5	2.2500	0.780	0.976
F6	0.3600	0.470	0.848
F7	0.0225	0.681	0.948
F8	0.0100	0.026	0.228
F9	0.1600	0.304	0.718
F10	0.3025	0.492	0.862
F11	0.0038	0.014	**0.165**
F12	0.1600	0.281	0.695
F13	0.0025	0.054	0.324
F14	0.0009	0.003	0.074
F15	0.000064	0.001	0.054

Table-5.3 Adjustability and Detectability (Calculation is based on true model and error covariance)

TAG	Error Variance	Adjustability	Detectability
F1	0.6456	0.558	0.897
F2	0.0578	0.082	0.397
F3	0.2536	0.358	0.767
F4	0.2545	0.791	**0.978**
F5	2.1273	0.778	0.975
F6	0.3531	0.468	0.847
F7	0.0225	0.705	0.956
F8	0.0066	0.018	0.186
F9	0.1593	0.299	0.713
F10	0.3122	0.498	0.865
F11	0.0044	0.016	**0.176**
F12	0.1656	0.289	0.703
F13	0.0021	0.046	0.299
F14	0.00079	0.002	0.069
F15	0.000065	0.002	0.055

Table 5.4 Adjustability and Detectability (Calculation is based on estimated model and error covariance in IPCA results)

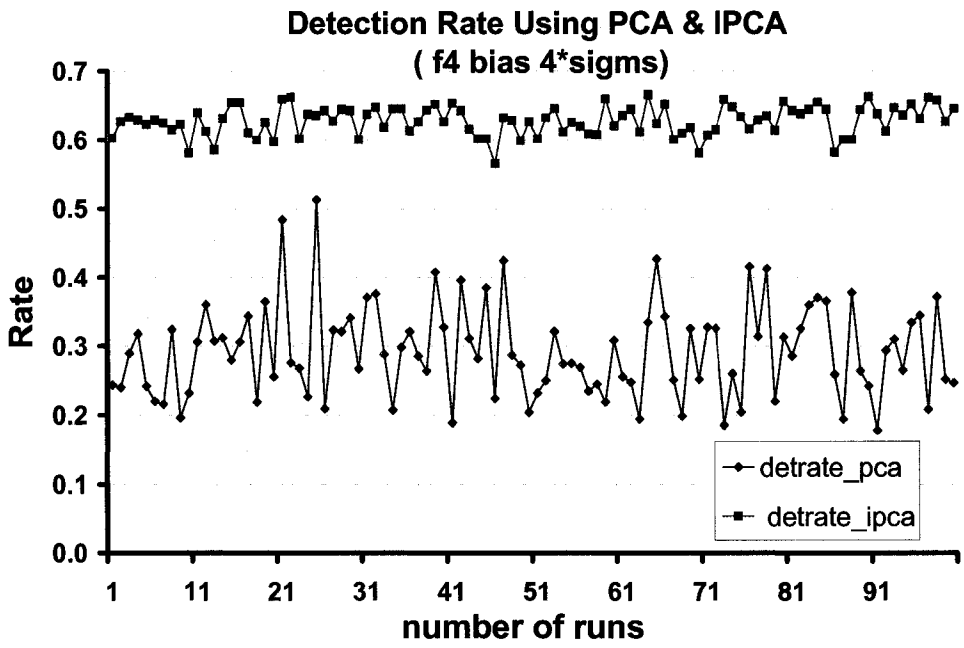


Figure 5.4 Detection rate using PCA-SPE and IPCA-SWR
(This assumes that the PCA model order is correctly selected as 7)

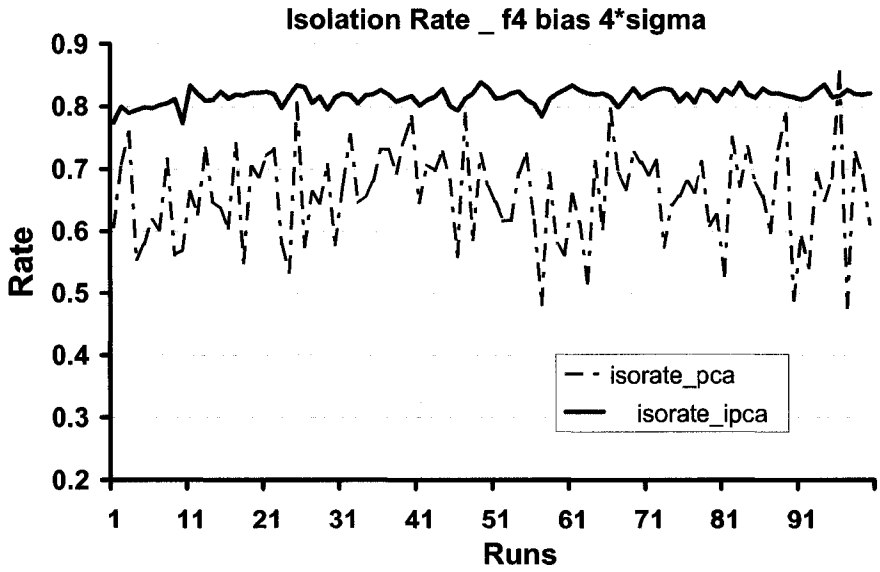


Figure 5.5 Isolation rate using PCA-SPE and IPCA-SWR-GLR
(This assumes that the PCA model order is correctly selected as 7)

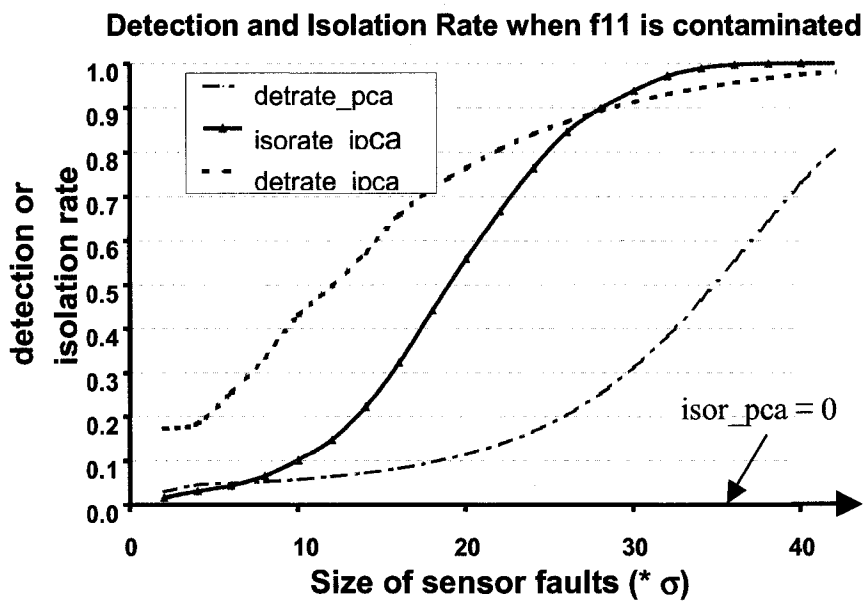


Figure 5.6 Detection and isolation rate using PCA-SPE and IPCA-SWR-GLR (This assumes that the PCA model order is correctly selected as 7)

Model Orders are Different

PCA cannot guarantee the correct selection of model order. In this case, without any hint from IPCA, PCA determines the model order to be 8 using the eigenvalue-one rule. Using different criteria, PCA may select different model orders. Under this consideration, Figure 5.7 gives a comparison of detection results using three approaches: PCA-SPE, PCA-SWR and IPCA-SWR, where PCA model order = 8 and IPCA model order = 7. In Figure 5.7 we can see that detection performance is much enhanced by using the IPCA-SWR method. PCA-SPE gives poor detection due to an incorrect model order selection, but PCA-SWR gives a much higher detection rate relative to that of the PCA-SPE method. This means that the SWR χ^2 test for detection is tolerant to a wrong selection of model order and is much superior than Q test, at least in this example. Similarly the isolation performances are also compared in Figure 5.8.

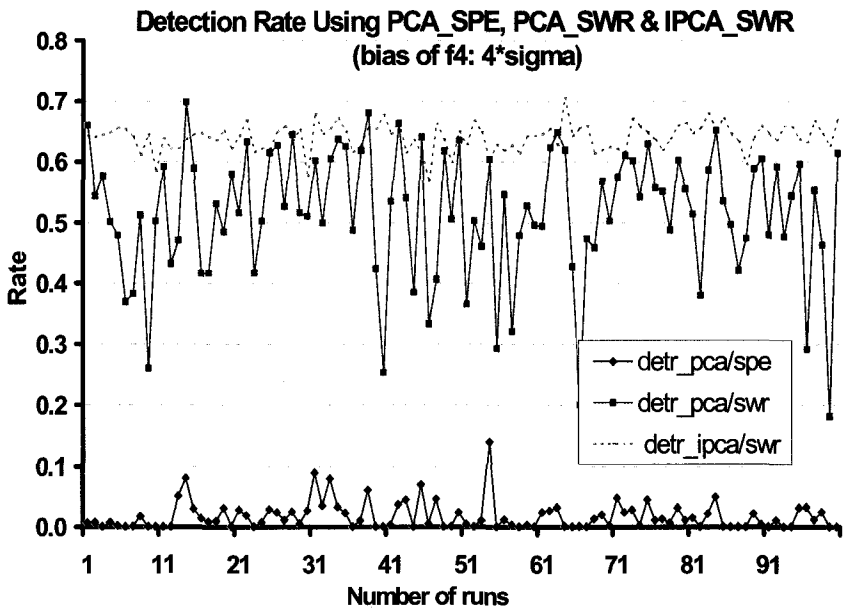


Figure 5.7 Detection rate using PCA-SPE, PCA-SWR and IPCA-SWR (PCA model order =8 and IPCA model order = 7)

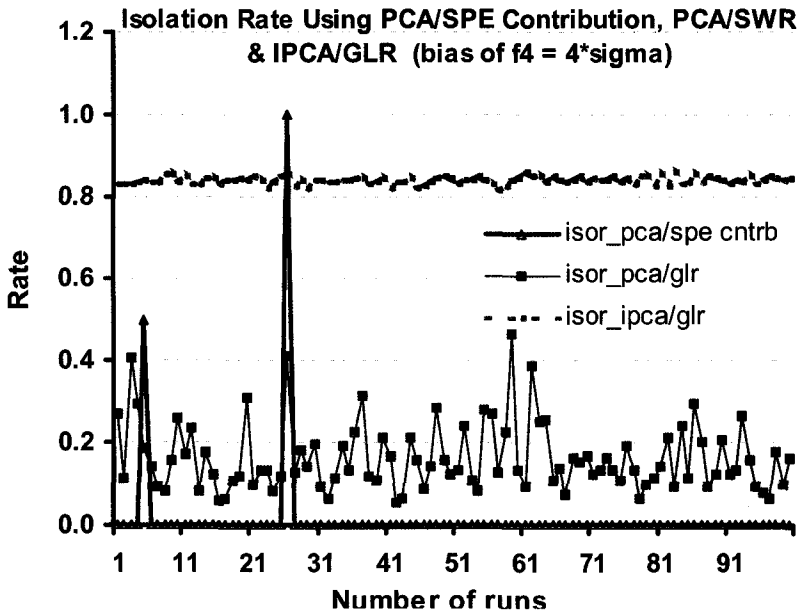


Figure 5.8 Isolation rate using PCA-SPE, PCA-SWR-GLR and IPCA-SWR-GLR (PCA model order =8 and IPCA model order = 7)

5.4 Conclusion

Although SPE and its contribution plot are traditionally used in FDI, SPE reduces the degrees of freedom of the test statistics. Commonly one-half to two-thirds of the diagnostic information (redundancy) in the data is lost using SPE. The SPE contribution plot method does not possess a convincing statistical basis and, in fact, sometimes gives an ambiguous diagnosis.

The proposed methods SWR and GLR enhance fault detection and isolation performance. A more accurate estimate of the constraint model can be expected using the IPCA algorithm and is helpful for FDI. The Mont-Carlo simulation results show the advantages of the proposed IPCA-SWR-GLR method for FDI purposes.

Sensor fault detectabilities of different measurements are related to system properties. Adding more sensors to the process in the process may enhance the system redundancy.

In practice, there is a certain amount of deviation from the Gaussian assumption that is essential for the statistical test: T^2 and SWR or SPE Q statistics; one solution is to train the thresholds from normal operation data.

Chapter 6

Comments on Practical Applications

6.1 Data Acquisition for PCA (or IPCA) Based Analysis

In the field of system identification, the data quality and the model structure knowledge are important for obtaining good process model estimations. The same thing happens to IPCA modeling: IPCA performance is dependent on data quality and data structure. There are a number of factors that affect IPCA performance.

6.1.1 Sample Size and Selection of Variables

Sample size

Application of the PCA method requires an adequate sample size to get a consistent result. Even though some techniques have reportedly provided a reasonably good solution from a limited amount of data (Rumantir, 1995), a reasonably large sample size is essential for estimating a reliable process model or prediction for a process. Fortunately, most application cases have enough data, and the recent growth in computational capacity has given us the power to deal with large sample sizes. When applying PCA, a larger sample results in a better fit. This is shown in Table 3.2.

Variable Selections

An appropriate selection of variables that are contained in the data matrix for IPCA analysis is critical. In real application we may have more than enough variables to be included in the IPCA modeling. Ignoring key variables results in poor model estimation, but including unrelated variables introduces extra calculation load and can also result in degradation of model estimation.

We know that IPCA estimates the error covariance matrix and therefore the scaling matrix $\mathbf{L}^{-1}$ according to the correlation structure of the data. In this thesis work, it is observed that if a variable has very limited correlation with all the others in the data matrix, then the estimation of the noise in this variable will not be an exact measure of its true noise, but is rather the summation of its true-signal excitation plus the true noise. In this case, we may find that the estimated SNR is close to one. So, in doing IPCA, whenever the estimated SNR in one variable is

close to one, we can conclude: (1) the excitation of this variable is not sufficient and should be increased to make it worthwhile to retain the variable in the data matrix for further analysis; or (2) the variable is intrinsically uncorrelated with others and should be removed from the current data matrix. If we know before hand the process knowledge, we can make our choice easily.

Tanaka and Mori (1997) gave the critira for variable selection for doing PCA. A windows package "VASPCA (Variable Selection in PCA)" was initially developed by Mori and Iisuka and has been converted to functions for use in general statistical packages, such as R and XploRe. The web-based software using the functions is also available.

It may be case that not all the useful variables are observable at each sampling instant. For instance, the multirate data is a kind of this. The focus of this thesis is mainly on steady state analysis of regularly sampled data. The case of multirate process data is beyond the scope of the present thesis.

The case of missing data in a steady-state problem is also beyond the range of the current topics in this thesis.

6.1.2 Input Probing or Process Excitation

To discuss the input probing and process excitation, we can not avoid talking about noise and the signal to noise ration (SNR).

Noise

Data is inherently contaminated by noise. The "noise" that is recognized as noise in PCA analysis may mean more than just random measurement errors. The origin and properties of noise can be of various types:

- Random measurement noise.
- High frequency part of true signal **X**, which may be recognized as "noise".
- Colored noise (auto-correlated noise) that is introduced by unmeasured disturbances.
- Spatially correlated noise.
- Noise correlated with signals.

Signal to Noise Ratio (SNR)

SNR is typically defined as the ratio of the signal power to the average noise power, and is typically measured in dBs. Here in this thesis, SNR is measured as the ratio of the standard deviation of the signal's excitation to the standard deviation of the noise. In the flow network example, two sets of fault-free data were generated with high and low SNR respectively (refer to table 6.1).

SNRs of measurement values may vary for different variables. Some are relatively high and others are relatively low. It has also been observed that PCA performs better if SNRs remain at a similar level as opposed to varying over a wide range. The adjustment can be made on SNRs via a proper scaling procedure (refer to the discussion on scaling methods in Chapter 2).

In the frequency domain, SNR varies with a shift in frequency bands. It is common in chemical processes that SNR is much higher within a slow frequency band than in the fast frequency band. This property makes it possible to apply a certain type of low-pass filter to reduce measurement error. An example is shown in Section 6.2.3.

Note that the measurements of different variables usually do not reach their highest SNR values at the same frequency band. Some measured inputs may have great power (i.e., large SNRs) at a very high frequency. However, this part of the information is commonly useless for steady state PCA modeling because most chemical processes perform as low-pass filters so that we cannot observe significant high-frequency responses on the output side.

	High SNR data				Low SNR data		
TAG	std_s	std_n	SNR	TAG	std_s	std_n	SNR
F1	13.51	0.8	16.89	F1	7.58	2.8	2.71
F2	10.94	0.25	43.76	F2	3.41	0.88	3.89
F3	5.72	0.5	11.43	F3	3.59	1.75	2.05
F4	5.65	0.5	11.31	F4	5.82	1.75	3.33
F5	5.58	1.5	3.72	F5	12.68	12.25	1.04
F6	12.47	0.6	20.78	F6	5.12	0.21	24.37
F7	10.67	0.15	71.11	F7	2.29	0.53	4.37
F8	3.02	0.1	30.18	F8	0.64	0.35	1.84
F9	14.58	0.4	36.45	F9	8.67	1.4	6.19
F10	14.85	0.55	27.00	F10	9.26	1.93	4.81
F11	2.9	0.06	46.77	F11	1.73	0.04	41.15
F12	16.52	0.4	41.29	F12	6.21	1.4	4.44
F13	10.66	0.05	213.23	F13	2.21	0.18	12.64
F14	7.01	0.03	233.76	F14	6.99	3.5	2.00
F15	0.24	0.01	29.86	F15	0.22	0.06	3.55

Table 6.1 SNR values in two sets of training data
(std_s: standard deviation of the signal's excitation;
std_n: standard deviation of the noise;
SNR: the ratio of the two.)

Higher SNR data contain more information. SNR affects the ICPA modeling results due to the reason stated below: Applying IPCA for model identification and error covariance estimation amounts to playing with the residual subspace of the whole matrix of the data set. Hence, how we determine or identify the residual subspace is crucial here; moreover, IPCA identifies this subspace according to the magnitude of the variance - those with the least variance will automatically be assigned to the residual space. It has been observed that PCA usually provides good estimate of constraint model from data with std (excitation) > 4×std (random noise).

6.2 Data Pre-Processing

Statistical techniques (such as PCA, CVA, PLS, etc.) for data mining are data-driven methods. Effectiveness of using these methods relies on the quality and properties of data, since abnormal data or error-contaminated data may severely distort the analysis results. In addition, the correct use of these statistical techniques implies that the data has to abide by certain assumptions. Therefore, data pre-processing is commonly used to facilitate these techniques.

6.2.1 Data Property

Stationarity

A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity means a flat looking time series, without trend, a constant autocorrelation structure over time and no periodic fluctuations.

To estimate a process model, we need enough excitation in the data with adequate SNR. In this context, data may move between different stationary operating points to have enough excitations. PCA generally does not require stationarity in data. However, we prefer that all the process data should fall in a relatively local operating region. In other words, data should move between different neighboring stationary points around a central operating condition. The reason is as stated bellow: The PCA method for modeling a process is to reveal the underlying correlation or the collinearity of the process data. This assumes that the process is linear. If there is any non-linearity in the process, PCA approximates the process property in a linear manner. This approximation is typically valid only over a limited operating region of the process, i.e., when the data is stationary over a limited local range.

For on-line monitoring, however, the operating region of the process may drift or change from one local condition away to another. This means that the on-line data range migrates temporally. Ordinary PCA is a batch-wise modeling approach in that the historical batch of data is acquired

from the process and is consequently used for PCA analysis. Hence, a question may be asked: should we update the PCA model on-line when new data are available? In fact, for on-line monitoring of a nonstationary nonlinear process, it is necessary to update the PCA model. Recursive PCA is one possible solution for this problem (Li *et al.*, 2000).

For a linear process, the shift in the data range does not affect the modeling result using PCA analysis. The data need not be stationary to get a good model and a good prediction, because the true model remains the same even if the operating point changes.

Sparse and Dense Data

This is concerned with the distribution of samples. We sometimes have much more data (observations) in a limited operating area and fewer observations in another operating area. As in the case of data stationarity, when systems show the property of non-linearity, direct use of sparse and dense data at the same time may cause a model-process mismatch in addition to a linearization mismatch. This is simply illustrated in Figure 6.1. The dash-dot line reveals the true relationship behind the noise-free data; the dotted line is the linear approximation of the true relationship over the range of observed data, which is the approximation we really mean to obtain; the solid line is the linear "curve" we estimate using PCA. We notice that the solid line leads to a significant model mismatch.

Good training data for linear model estimation should possess the quality that all data clusters from different areas have a similar leverage on the identification result. If we have *a priori* knowledge that the underlying true model is linear (the dashed line is an absolutely straight line), then we would not be concerned with this problem.

White Noise, Single-Frequency and Random-Binary Signal

The spectrum of data excitation is evenly distributed over all frequencies when the signal is white. However, it is usually true that only sensor errors rather than the signal excitation are white noise. In practice, a common scenario is that most signals are colored. Extreme examples are single frequency signals and random binary signals. The spectrum of the data in the flow network example is shown in Figure 6.2.

Because most signals are colored and sensor errors (noise) are white, a filter can usually be designed for data cleaning, which is intended to result in a certain degree of enhancement of SNR properties. More discussion on this point will be given in Section 6.2.3.

Random binary signals approximate white noise. Band limit random binary signal is white noise through a band-pass filter and then converted into a binary sequence.

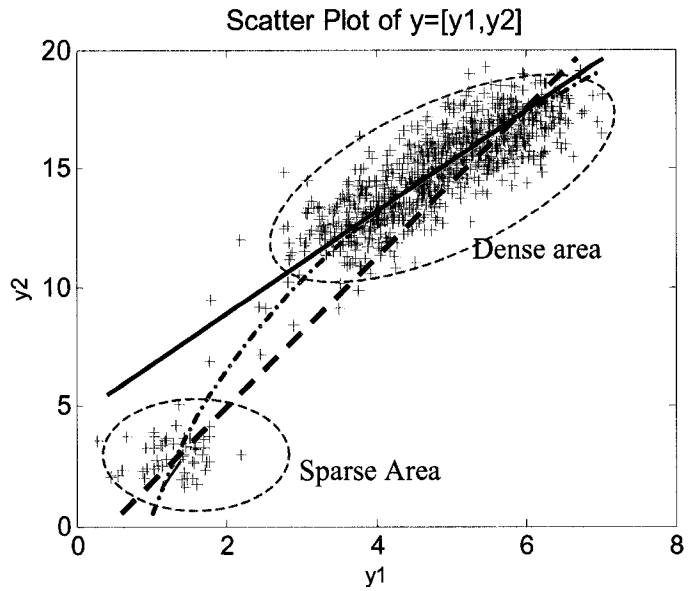


Figure 6.1 Sparse and dense data from a nonlinear system causes additional model-process mismatch

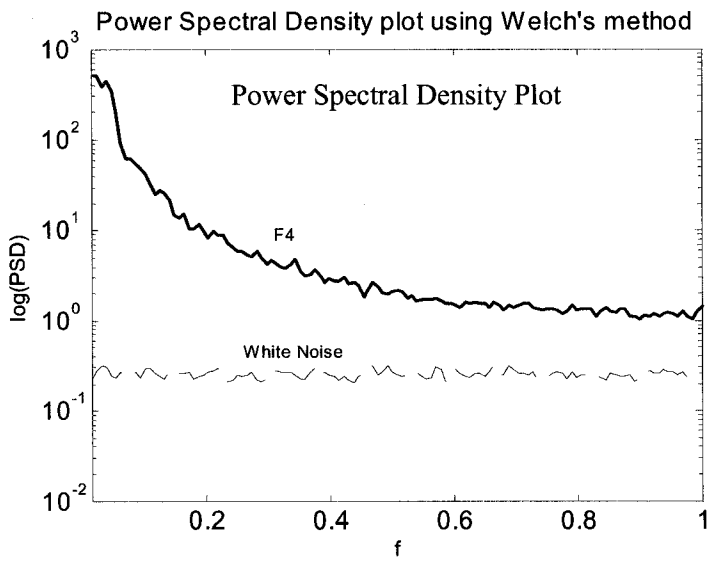


Figure 6.2 Power Spectrum of the signal in F4 shows its color property (sample size =20000)

Steady State Data and Transient Data

The system in steady state is at thermodynamic equilibrium. In other words, if a system is in steady state, the mass, heat, and momentum balance at each local position in the system are in equilibrium all the time. If any extraneous force exerts on this system, the equilibrium will no longer hold and the system will start to move (change with time) until it reaches another equilibrium after a certain amount of time. The process of the system moving from one steady state to another is known as a transient condition.

In the real world, a system is hardly ever at steady state, and essentially is in transient condition. In this respect, the analysis of steady-state data interspersed with transient states can be defined as *quasi steady state* data. The data will be highly auto-correlated if there is significant dynamic feature in the data. If the data available for analysis is highly auto-correlated, then special care should be paid in collecting and analyzing raw data from a process.

Steady-state data is useful for data reconciliation and process optimization because steady-state data reveals important properties of a process and, at the same time, does not include the complexity and uncertainty in the dynamic data. However, sometimes we focus on the dynamic properties of a process, for instance, when we are concerned with system identification and automatic control problems. If this is the case, then we prefer that data retain the most transient part of the system. In PCA modeling, quasi-steady state data results in the steady-state process model, but properly collected transient data is expected to result in a dynamic process model. How to collect the data does affect the PCA analysis results.

Given a data set in hand, it is critical to tell which segments of the data represent the quasi-steady state of the process and which segments represent the transient condition. Wavelet transform (WT) is used to detect steady state for continuous process. Because the first-order WT of a time series is proportional to its first derivative, the corresponding wavelet modulus measures the variation in the underlying process trends. The process is said to be in steady state when the modulus is equal or close to 0.

Spatial Correlation and Temporal Correlation

Spatial correlation, typically used in the fields of image analysis and climate modeling, describes the correlation between signals at different points in space. In chemical processes, spatial correlation is the correlation between different variables involved in the data set. For steady state model identification, PCA requires quasi-stationary data to determine the relationships among different variables, i.e., the *spatial correlation* in the data.

Temporal correlation describes the correlation between signals observed at different moments in time. This correlation for the same variable is called *auto-correlation* and for different variables

(or different points in space) is called *cross-correlation.* Dynamic PCA extracts both the spatial and temporal correlation among observed variables.

Gaussian and Non-Gaussian Time Series

Given a set of raw data, we can use a *Q-Q plot* to evaluate the normality of the *Univariate Marginal Distribution* for each individual variable. A *chi-squared plot* is applied to check multivariate normality (Johnson, 1998). Appropriate transformations such as maximal power transformation (Box & Cox, 1964) can make the data more normal, but such nonlinear transformations may also distort the underlying model in PCA analysis. So, nonlinear transformation should be carefully used before PCA modeling. Choudhury and Shah (2003) suggest a Non-Gaussianity Index (NGI) for evaluating normality based on the bicoherence of signal from the higher order statistical point of view. However, in this thesis, PCA analysis is based on the *first and second order statistics* (mean, variance, autocorrelation, power spectrum) of the data. The *Q-Q* and *chi-squared plots* are sufficient to evaluate the normality of the data. Another straightforward method to get a rough idea of the normality of univariate data is to plot a *histogram* of the data.

The *Q-Q* plots in Figure 6.3 evaluate the normality of simulated data for F2 in the flow network process (Figure 3.1). The left hand-side plot shows that the measurement noise in F2 is normal while the right hand-side plot shows that the observed data F2 (signal + noise) is not normal. In fact, the main part of F2 consists of band-limited random binary excitation.

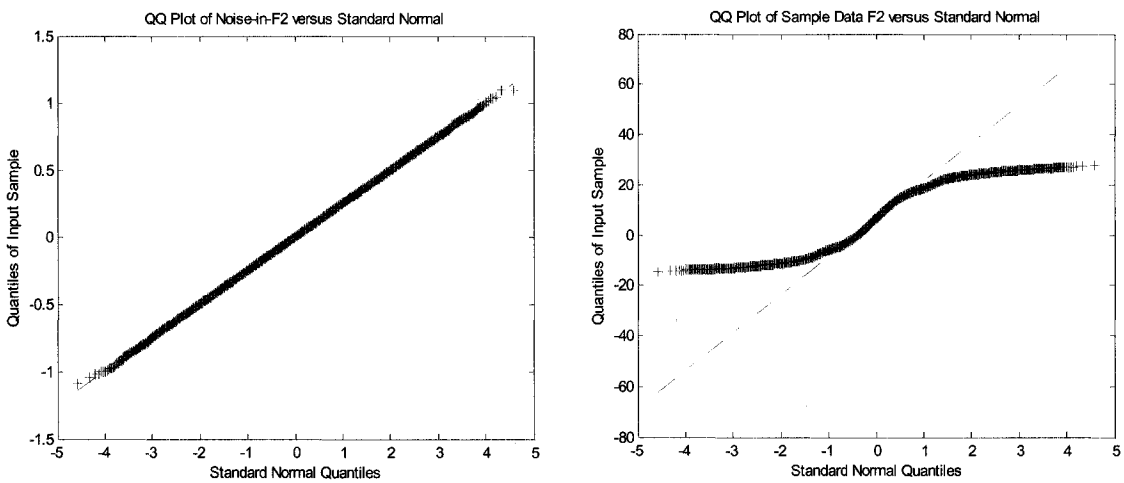


Figure 6.3 *Q-Q* plots to evaluate the normality of data

6.2.2 Outlier Detection and Elimination

What is an outlier?

An outlier is one observation that appears to deviate markedly from other members of the sample in which it occurs (Grubbs, 1969). Hawkins (1980) also gave a similar definition: "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism." Outliers can be classified into four classes as shown below (Hair *et al.*, 1998):

- Procedural error, such as a data entry error or a mistake in coding. These outliers should be identified in the data cleaning stage.
- Extraordinary event, which then is an explanation for the uniqueness of the observation.
- Abnormal observations that the researcher may not be able to rationalize.
- From a multivariate point of view, outliers may be observations that fall within the ordinary range of values on each of the univariate variables but are unique in their combination of values across the variables.

Outlier Detection and Elimination

Many statistical techniques have been proposed to detect outliers and comprehensive texts on this topic are those by Hawkins (1980), Barnet and Lewis (1994). The univariate perspective for identifying outliers examines the distribution of observations and selects as outliers those cases falling at the outer ranges of the distribution. The detection method is relatively straightforward and the primary issue is to establish the threshold for designation of an outlier. One heuristic criterion defines a value more than three standard deviations away from the mean as an outlier. Scatter plots can be used for checking pairs of variables jointly to detect a case that falls significantly outside the ellipsoid formed by other observations as an outlier. An ellipsoid of a confidence limit can be applied to facilitate identification of the outliers.

For multivariate or highly structured data, more sophisticated methods are proposed. The statistical methods that have been developed for this purpose include graphical and pictorial methods (Kleiner and Hartigan, 1981), principal components-based methods (Hawkins, 1974). Neural network based outlier detection has also been developed (Liu *et al.*, 1998). In applying these methods, we want to avoid two things: the blind removal of outliers, which may result in loss of information and therefore often too simplistic a model, and an over-fitted model, which may result in poor performance during cross validation, i.e., it will model random noise. On the attempt to balance these two things, researchers such as DeBoer and Feltkamp (2000) suggested robust techniques for outlier detection and elimination. Liu *et al.* (2002) have given a critique and a solution for determining which kinds of outliers should be deleted and which should be retained for knowledge extraction.

6.2.3 Moving Average Filtering

EWMA Filter

Filters have been widely used for reducing random noise in process data. Slow sampling rates generate low-frequency signals from high-frequency signals, in which case signal aliasing may occur. To avoid aliasing, process data is pre-filtered in an analog device before sampling. Digital fiters, such as the exponentially weighted moving average (EWMA) filter, are applied to the sampled data to further reduce high-frequency noise. EWMA is also used for obtaining data for univariate statistical control charts in process monitoring.

In the flow network example (Figure 3.1), EWMA was used to reduce measurement noise. EWMA filters take the form of a first-order process, just like a surge tank. Its expression is as follows:

$$y_t^f = (1-\alpha) y_{t-1}^f + \alpha y_t \qquad \ldots\ldots\ldots (6.1a)$$

where: y_t is the observation at current time t, y_t^f is the filter output at time t, and α is the filter parameter.

Effect of EWMA Filters on White Noise

In the forthcoming dicussions we need to know how EWMA filters change the covariance of white noise. In the form of the *infinite impulse response* (IIR), an EWMA filter can be expressed as:

$$y_t^f = (1-\beta)\left(1+\beta q^{-1}+\beta^2 q^{-2}+\cdots\right) y_t \qquad \ldots\ldots\ldots (6.1b)$$

where $\beta = 1-\alpha$. In general, if a set of multivariate data

$$y_t = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_t$$

and the same EWMA filter is applied to it, then we have:

$$y_t^f = \left(\mathbf{A}_0 + \mathbf{A}_1 q^{-1} + \mathbf{A}_2 q^{-2} + \ldots..\right) y_t = \Xi_{0,ewma} y_t \qquad \ldots\ldots\ldots (6.2a)$$

where $\mathbf{A}_i$ is diagonal and

$$\mathbf{A}_i = (1-\beta)\beta^i \bullet \mathbf{I}, \quad i = 1, 2, \cdots \infty \qquad \cdots\cdots\cdots (6.2b)$$

Different from equation 6.1b, y_t in equation 6.2a represents no longer a univariate variable but a multivariate variable. Equation 6.2a is a linear transformation of y_t .

Let us look at the white noise e_t in this variable. Because $y_t = x_t + e_t$, and $e_t \sim N(0, \Sigma_e)$, from equation 6.2a we have

$$e_t^f = \left(\mathbf{A}_0 + \mathbf{A}_1 q^{-1} + \mathbf{A}_2 q^{-2} + ...\right) e_t$$

Given e_t is iid normal, we have

$$\Sigma_{ef} = \left(\mathbf{A}_0 \Sigma_e \mathbf{A}_0^T + \mathbf{A}_1 \Sigma_e \mathbf{A}_1^T + \mathbf{A}_2 \Sigma_e \mathbf{A}_2^T + \cdots\right) \qquad \cdots\cdots\cdots (6.3a)$$

Combining with equation 6.2b, this can be simplified as

$$\begin{aligned} \Sigma_{ef} &= (1-\beta)^2 \left(1 + \beta^2 + \beta^4 + \cdots\right) \Sigma_e \qquad \cdots\cdots\cdots (6.3b) \\ &= \frac{(1-\beta)\left(1-\beta^{2(n+1)}\right)}{1+\beta} \Sigma_e \\ &= \frac{1-\beta}{1+\beta} \Sigma_e, \quad (\text{when } n \to \infty) \end{aligned}$$

This result will prove to be useful later. From this equation, we can see how an EWMA filter reduces the variance of white noise. A selection of the filter parameter $\beta = 1 - \alpha$ can determine how an EWMA filter changes the covariance of a white noise.

Notice that equation 6.3b is derived for white noise. If a vector of colored multivariate signals passes an EWMA filter $\Xi_{0,\,ewma}$, then the change of its covariance will not be simply estimated by equation 6.3b. However, we can do that by following steps:

- Find a numerical estimate of the filter Ω_j for each colored signal x_j (j = 1~n) so that:
  $x_j = \Omega_j w_j$, where w_j is a white noise.
- Redefine a new filter for each of them as $\Xi_{j,\,new} = \Xi_{0,ewma} \Omega_j$.
- Develop IIR expressions for Ω_j and $\Xi_{j,\,new}$ (similar to equation 6.1b).

- Estimate the covariance matrix of the whitened signals w_j by solving equation 6.3a for Σ_e, and notice that $\mathbf{A}_i$ is obtained from the group of Ω_j ($j = 1 \sim n$)
- Obtain the covariance of filtered signals using equation 6.3a again, and notice that $\mathbf{A}_i$ is obtained from the group of $\Xi_{j,new}$ ($j = 1 \sim n$)

Most chemical process variables are color. Given a process variable $y_j = x_j + e_j$, where $x_j = \Omega_j w_j$ is the true colored signal and e_j is the white measurement noise. When an EWMA filter $\Xi_{0,ewma}$ is applied to y_j, then:

$$y_j^f = \Xi_{0,ewma}\left(x_j + e_j\right) = \Xi_{0,ewma}\Omega_j w_j + \Xi_{0,ewma}e_j = \Xi_{j,new}w_j + \Xi_{0,ewma}e_j \quad \ldots\ldots \quad (6.4)$$

This means that two different filters are applied to w_j and e_j. In this context, supposing the measurement errors (noises) are white, carefully selected EWMA filter can enhance SNR in the data. Figure 6.4 shows the selectivity of parameter α in upgrading the SNR of the data from the simulated flow network. We find that the selectivity of parameter α varies for different variables. Ideally, the observations on all variables should have the maximum sufficient excitation power over a common range of frequencies. However, this desirable scenario seldom exists. We cannot tailor the filter parameters to the various respective variables because this results in applying different filters at the same time, which is not a linear operation. Before using PCA, we can only apply a common filter $\Xi_{0,ewma}$ to all variables at the same time.

In Figure 6.4 we can see that $\alpha = 0.05$ gives the greatest enhancement of SNR. However this α value does not guarantee, as found in the flow network example, a good result for PCA modeling. This phenomenon is understandable: if α is so small as 0.05 (in other words, the filter is very strong), the data quality (SNR) will be optimally enhanced but at the cost of losing too much amount of signal power.

Even though low pass filters such as EWMA may reduce data noise, they should be applied with caution. After all, filters introduce time delay and extra dynamics to the data, i.e., add temporal correlations to the data. These temporal correlations will degrade the statistical assumptions for building thresholds in SPE or T^2 monitoring charts.

As mentioned before, a chemical process is like a filter. If the process data is transient or dynamic, we should apply dynamic PCA on this data. As just mentioned, the temporal correlations in the data will make it complecated to compute the thresholds in SPE or T^2 monitoring charts. In such a case, we use either heuristics or more complex statistical techniques to monitor process performance. We can also use wavelet transformation to apply MSPCA to handle this difficulty (Bakshi, 1998; Luo *et al.*, 1999; Misra, 2002).

Applying IPCA to Filtered Data

For low SNR data, we can apply an EWMA filter first and then IPCA, but we should tune the IPCA algorithm under the assumption that all errors represent white noise.

As shown in figure 6.4, a proper EWMA filter may enhance the SNR of the observed data especially when the measurement errors have dominant power in a high-frequency band (which is often the case in chemical processes). Our study shows that a better result is obtained when PCA is applied to filtered data than to non-filtered data. If the SNR in the data is large, this enhancement is not significant. In that case, filtering is not required.

We can apply PCA to filtered data because a filter, as a linear operation in temporal space, will not affect the spatial model. The only difference is the error covariance matrix as shown by equation 6.4. We cannot apply IPCA directly to filtered data because equation 2.17 requires errors to be iid, and a filter introduces temporal correlation to the data. We should use the original data to estimate Σ_e by equation 2.17.

The algorithm of IPCA for EWMA-filtered data is given in figure 6.5. In the figure the scaling matrix $\tilde{\Sigma}_{y^f}$ has the same diagonal elements as Σ_{y^f} and zero off-diagonal elements.

Exponentially Weighted Moving Covariance (EWMC)

EWMC is an EWMA operation on the data covariance matrix. Equation 6.5 shows how the moving covariance Σ_t^f is updated:

$$\Sigma_t^f = \beta\Sigma_{t-1}^f + (1-\beta)\Sigma_t \qquad \ldots\ldots\ldots \quad (6.5)$$

where $\beta\,(0 \le \beta \le 1)$ is the weighting. Here t may represent a time instant, or instead, represent a batch-wise window. For example, a new batch data ($t = i+1$) is first compared to a PCA model based on the moving covariance $\Sigma_{t=i}^f$ that only depends on past data. Then, a new updated moving covariance will be calculated if necessary. The procedure is repeated recursively to obtain an on-line PCA model. Similar to the discussion on EWMA filter, the filter parameter β should be carefully selected.

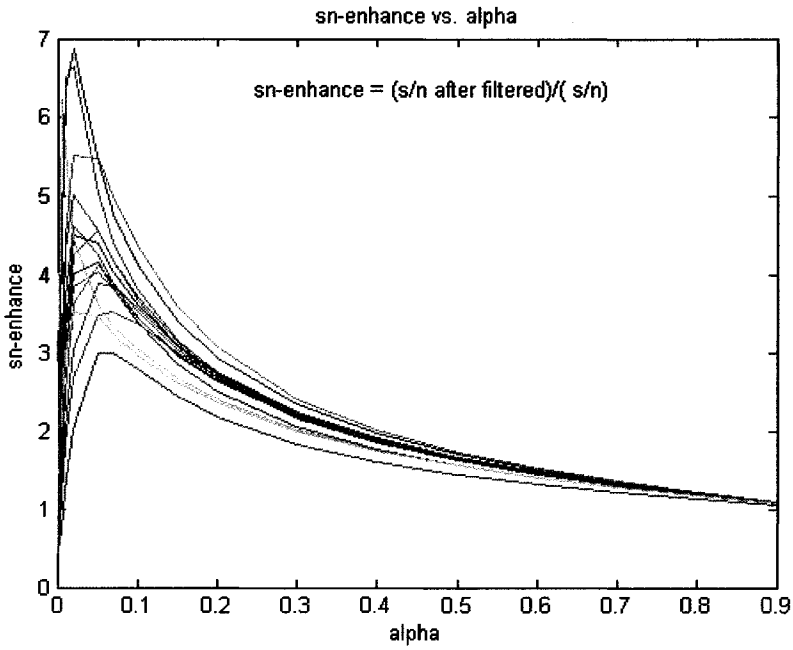


Figure 6.4 EWMA enhanced SNR of the flow network data

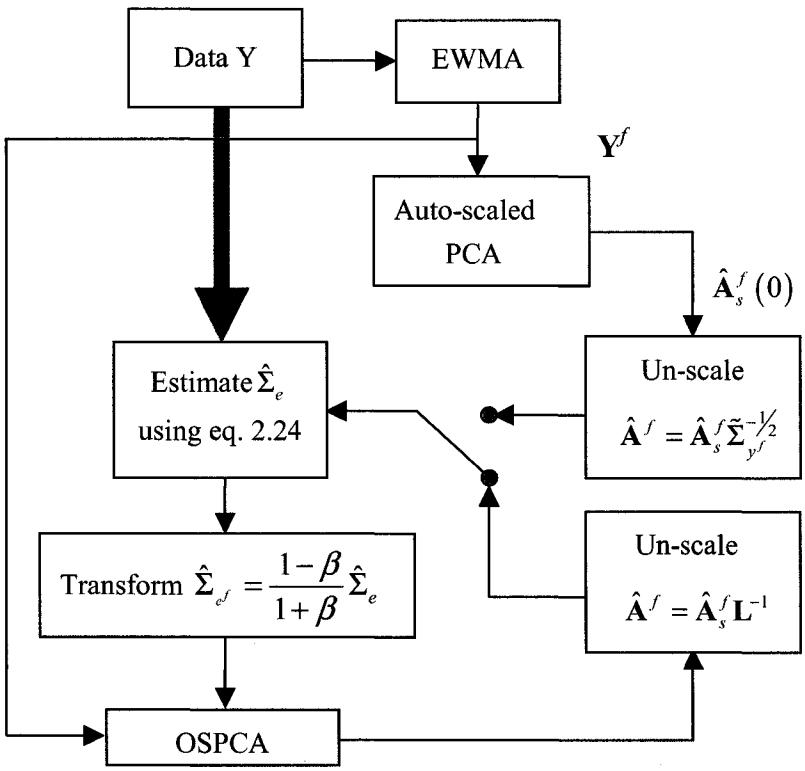


Figure 6.5 Procedure of applying IPCA to filtered data

6.3 Comments on Application of PCA/IPCA for Model Identification

6.3.1 Apply IPCA to Dynamic Process Data

In dynamic cases, the data matrix structure may change the model order recognized by checking the number of unity eigenvalues when IPCA estimates a model. For instance, if we have a data set defined as $\mathbf{Y}=\begin{bmatrix} y_t & y_{t-1} & u_{1,t} & u_{1,t-1} & u_{2,t} & u_{2,t-1} \end{bmatrix}$, and the sampling rate is fast, we tend to numerically add the following correlations to the data:

$$\begin{aligned} y_t - y_{t-1} &\approx 0 \\ u_{1,t} - u_{1,t-1} &\approx 0 \\ u_{2,t} - u_{2,t-1} &\approx 0 \end{aligned} \qquad \ldots\ldots\ldots (6.6)$$

Equation holds when the time series are not white; although the process inputs can be fast and very dynamic (unstationary), equation 6.6 inevitably comes true unless the sampling rate is slow. This is because most chemical processes behave like low-pass filters so that the variables exhibit strong autocorrelation. To handle this problem, we have two choices:

(1) Choose slow sampling rate;
(2) Perform a linear transformation on the data set as shown in equation 6.7.

$$\mathbf{Y}_{new}=\mathbf{Y}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}=\mathbf{YM}=\begin{bmatrix} y_t & \Delta y_t & u_{1,t} & \Delta u_{1,t} & u_{2,t} & \Delta u_{2,t} \end{bmatrix} \qquad \ldots\ldots (6.7)$$

If we apply IPCA to the new constructed data set in equation 6.7, then the extra confusing relationships in equation 6.6 are avoided, the true steady-state model and the fist-order dynamic model can be correctly identified.

This linear transformation is only recommended for the case of first order systems. If a process exhibits second order dynamics, the problem becomes more complicated.

Recall that IPCA needs the condition:

$$\frac{m(m+1)}{2} \geq n \qquad \ldots\ldots\ldots (6.8)$$

Where m is the number of constraints, n is the number of unknown elements in the error covariance matrix Σ_e. If Σ_e is diagonal, this number is commonly equal to the number of variables in the data matrix. However, for Y_{new} in equation 6.7, n can be reduced to 3 because the noise in Δy_t can be represented by the noise in y_t and so on. If the noise in y_t is white, then we have: $\text{var}(e_{\Delta y_t}) = 2\,\text{var}(e_{y_t})$.

6.3.2 Using IPCA Outputs to Facilitate PCA Modeling for On-Line Implementation

In practice, there is no guarantee that equation 6.8 will always be satisfied to make IPCA applicable, in this case PCA is useful. When equation 6.8 is satisfied, IPCA model can be time consuming, making it hard to use for on-line recursive implementation. The possible solution is generally to perform IPCA modeling once and then apply OSPCA modeling procedure in combination with the given knowledge of model order and the estimation of the error covariance. The OSPCA can provide fairly good performance given that the model order selection is correct and the error covariance is roughly correct.

6.3.3 Dealing with Color Noise

PCA apparently does not require the noise distribution to be normal for PCA modeling, but implicitly requires it for Q statistics. IPCA requires noise to be white in drawing the maximum likelihood function for the iterative implementation.

However, it has been observed that IPCA works well for noise with weak auto-correlation (for instance, white noise is filtered by EWMA with $\alpha = 0.7$ in equation 6.1a). If noise is highly auto-correlated, then Multi-Scale PCA is preferable. Issues of this kind are discussed in detail in the literature (e.g., Bakshi, 1988).

6.3.4 Dealing with Perfect Measurements

We rarely observe a perfect measurement (unless the number of people or things of that kind) of a process variable. However, some variables can be measured at an extremely high level of accuracy. These kinds of measurements make the error covariance matrix Σ_e singular, i.e., at least one of its diagonal elements is almost zero. If we use IPCA, the estimate Σ_e will be ill-conditioned and the solution will be unstable. If this is the case, we will have trouble to calculate the inverse term in equation 2.17.

A simple way to solve this problem is to add errors to the perfect measurements so that the new constructed data have a reasonable SNR.

Given $\begin{bmatrix} y_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ 0 \end{bmatrix}$, if we add small errors δ to the perfect measurement x_2, then the new "observed" data can be represented as:

$$\tilde{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ \delta \end{bmatrix}$$

Then the error covariance matrix takes the form:

$$\Sigma_e = \begin{bmatrix} \Sigma_{e_1} & \mathbf{0} \\ \mathbf{0} & [\mathrm{diag}(\delta)]^2 \end{bmatrix}$$

We can then perform IPCA on data $\tilde{\mathbf{Y}} = [\tilde{y}_1 \quad \tilde{y}_2 \quad \tilde{y}_3 \quad \cdots \quad \tilde{y}_N]$ and estimate the unknown elements of Σ_e.

6.4 Comments on Application of PCA/IPCA for Fault Diagnosis

6.4.1 Applying SWR/GLR Strategy for Fault Detection and Isolation

No matter what model is obtained from PCA or IPCA, SWR and GLR should be used in fault detection and isolation instead of using SPE's Q statistics and SPE contribution plots. For a large-scale problem where there are a great number of variables included in the data matrix, PCA is useful because it does not require that equation 6.8 hold true. As noted in this thesis, the number of PCs retained in the principal component subspace (PCS) is important for efficient FDI. The number of PCs can be determined by IPCA (if applicable) or by training the model for the highly sensitive response of a set of given faults (if the properties of these faults are previously known).

6.4.2 Applying Filters if Necessary

Using a filter properly is almost an art in itself. Filtering has three main functions: (1) Enhances SNR ratio if true signal is slow and noise is fast or white; (2) Reduces false alarms and detects

minor sensor biases; (3) Gets rid of the undesirable effect of autocorrelation (in doing statistical analysis) by using a wavelet filter.

If possible, we should avoid filtering the data before our analysis. However, in the case that the data have a very small SNR, it becomes impossible to observe the process clearly over the whole frequency altogether. Therefore, EWMA filtering may enhance the SNR, as shown earlier. The drawback is that a moving average filter introduces autocorrelations to the signal and colors white noise. Therefore, if we apply a moving average filter in doing PCA or IPCA, we need to deal with the color noise. Great care should be paid to make clear if the autocorrelation that a moving average (such as EWMA) filter introduces to the data in the analysis is critical or not. The PCA method is robust to autocorrelation, but the threshold for Q statistics or the SWR test should be adjusted accordingly. However, IPCA is not very robust to color noise according observations in doing this thesis work.

Wold (1994) discussed the use of EWMA filters in combination with PCA and PLS. In fault detection, **we can apply a filter just before plotting the results** (SPE, SWR, T^2) to reduce false alarms and, at the same time, to enhance the sensitivity to detecting minor faults. The price to pay is the time delay in fault detection.

6.4.3 Multiple Gross Error and Process Fault Detection and Diagnosis

Process abnormalities may result from system faults other than sensor faults, for instance, process leaking, control valve stiction, fouling of heat exchangers, poisoning of catalysts, and so on. GLR can easily detect simple sensor faults and even process leaks. However, for the detection and isolation of multiple gross errors and/or complex system faults, things become complex. A multiple sensor fault is a combination of a group of simple sensor faults. A multiple gross error is a combination of a group of simple gross errors (including leaks). When facing a multiple gross error problem, we need to take into account the equivalency of gross errors when choosing among the suspected combinations of simple gross errors (Jiang *et al.*, 1999). Rosenberg *et al.* (1987, 1999) and Sanchez (1999) have designed various strategies for simultaneous identification and estimation of sensor biases and process leaks.

Process fault detection and isolation is even more complicated than multiple gross error problems. For instance, a vale stiction can lead to nonlinear process performance. In that case, many variables can be affected in different ways. The detection and especially the isolation of a system fault are very difficult without any a priori knowledge of the fault in question.

If a process fault results in a consistent pattern of fault signatures in PCS and/or RS, we can try to obtain this pattern by means of training. Through training, we can assemble a fault bank capable of recognizing process faults in future monitoring. It is also beneficial to use process knowledge in an attempt to carry out complex fault diagnosis. Detailed discussion of this point is beyond the scope of this thesis.

6.5 Other Considerations When Applying PCA or IPCA

There are many practical issues when people come to apply PCA or IPCA to real process operation scenarios. These include:

- How do we recognize and deal with data preprocessing impacts such as data averaging, data compression, truncation, and quantization?
- How do we deal with missing data or multiple sampling rates?
- How do we get rid of fault propagation problems when the system is under closed loop controls (an area in need of more inside study)?
- How do we recognize the existance and estimate non-diagonal elements in the error covariance matrix?
- How do we solve alignment problems in a batch process?

Some of these topics – such as PCA application in batch processes (Meng, 2000; Salvador *et al.*, 2003) and dealing with missing data problems (Nelson, 1996) – have been extensively discussed in the literature. However, there is still a lot of room for us to gain further knowledge on these problems both through research efforts and practical applications.

Chapter 7

Conclusions and Future Work

This thesis has examined the basic application of multivariable statistical tools for process engineering concerns – linear model identification, data filtering, and sensor fault detection and diagnosis. The main focus of this thesis is put on a novel PCA approach, IPCA, for process model identification and its application in different FDI schemes. The thesis goes through three parts: First, it provides the introductory background on this thesis subject as well as a basic preparatory discussion on the data qualities that the author believes to be important. Second, the thesis introduces an improvement on the current PCA method, proceeding in a series of steps from PCA to MLPCA, OSPCA, and then to IPCA. The new method was verified both by comparing similarities to MLPCA and by simulation studies. A geometric explanation has also been given to help understand the necessity for and the advantage of using IPCA. A comparison between OLS and PCA as regression tools, as well as a comparison between DR, IPCA, and PCA as data filters, has also been made. In the last, the thesis focuses on FDI problems and suggests IPCA-SWR-GLR, a novel combinatorial scheme for sensor fault detection and isolation. The power of this new scheme was verified both by comparative criticisms of the traditional SPE-based methods and by simulation studies.

7.1 Contributions of This Thesis

This thesis study offers the following contributions:

- Data qualities are discussed via a flow network simulation example.
- The weakness of PCA methods both in model ID and sensor FDI is presented through careful studies.
- IPCA, a novel approach for process modeling, is presented, showing its great superiority over PCA. The improvement made by this new method was verified both by equations and through simulation examples.
- An effective combinatorial FDI scheme, i.e., IPCA-SWR-GLR, is proposed for sensor fault detection and isolation. The new scheme was verified to be more powerful than the current PCA-based methods via Mont Carlo Simulations.

7.2 Concluding Summary and Directions for Future Research

On the basis of the research carried out for this thesis, we can come to the following conclusions.

- IPCA provides an optimal solution for decoupling signal from noise and determining the model order. It represents a great improvement over conventional PCA.
- Using IPCA, we can simultaneously obtain the process model, the measurement error covariance, and determine model order.
- In doing IPCA, scaling parameters are optimally and automatically selected, that is, IPCA is a scale invariant technique.
- Given either the true process model or the true error covariance, IPCA becomes MLE. In that case, the IPCA filter is equivalent to DR.
- Significant improvement can be made to PCA-based modeling and it's consequent FDI scheme PCA-SWR-GLR, with the *a prior* condition that model order has been correctly selected, at least for the detailed (Monte-Carlo) example considered here.
- SWR and GLR are more effective in the detection and diagnosis of sensor faults than are SPE-based methods (Q statistics and SPE contribution plot), at least for the detailed (Monte-Carlo) example considered here.

The work carried out for this thesis raises a number of questions and provides some directions for future work. In particular, the following topics merit consideration from researchers.

- The optimal selection of variables to be included in the analysis. Including more variables retains more information, however, it also introduces more uncertainty and noise, and increases computational load. We need to answer such questions as: should we retain a variable that we know has strong links to properties of interest when the measurement of this variable is very noisy?
- The extension of the optimal scaling idea and the iterative approach for error covariance estimation to PLS and CVA.
- The extension to subspace-based dynamic model identification. Further work might include derivation and verification of the new approach, which could be followed by a comparison of the new approach and the PEM method.
- Deriving the confidence intervals for estimated parameters may provide us with an inference as to how reliable the estimation results are. So it is worth a try.
- The error covariance matrix is only valid for white noise. Unfortunately, real processes are subject to unmeasured disturbances that will automatically be recognized as color noise. If this is the case, a spectrum description is a fair measure of noise and

unmeasured disturbances. It would be worthwhile to do more work on noise spectrum estimation.

- The schemes for complex fault detection and diagnosis need to be further studied. In doing this, we can provide an answer as to how to combine PCS and RS for this purpose. Yue and Qin (2001) proposed a combined index for fault detection, but is there a better way of handling both detection and isolation?
- For the dynamic case with closed control loops, it might be possible to borrow ideas from the *Kalman Filter* or other possible model prediction techniques for the on-line score trajectory prediction. The on-line score trajectory will help reveal important information for complex fault FDI.
- If we already know the fault features or we have enough faulty data for training purposes, it would be a rewarding try for us to use the clustering techniques or pattern recognition techniques jointly with the IPCA-SWR-GLR scheme to detect complex faults.
- Quantitative and qualitative analysis of model mismatch would be helpful to derive a robust method for on-line process monitoring. The objective is to reduce false alarms with tolerance of model or process drift, maintaining sufficient sensitivity in fault detection. Efforts need to be made to achieve this goal.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **19**, 716-723.

Almasy, G. A., and T. Sztano (1975). Checking and Correction of Measurements on the Basis of Linear System Model. *Problems of Control and Information Theory*, **4**, 57-69.

Bagajewicz, M. J., and Q. Jiang (1998). Gross Error Modeling and Detection in Plant Linear Dynamic Reconciliation. *Computers Chem. Engng.*, **22**(12),1789-1809.

Bagajewicz, M., Q. Jiang and M. Sánchez (1999). Performance Evaluation of PCA Tests for Multiple Gross Error Identification. *Computers and Chemical Engineering*, **23**, Supp., 589-592.

Bagajewicz, M., Q. Jiang and M. Sánchez (2000). Performance Evaluation of PCA Tests in Serial Elimination Strategies for Gross Error Identification. *Chemical Engineering Communications*, **183**, 119-139.

Bagajewicz, M. and D. K. Rollins (2003). On the Consistency of The Measurement Test and GLR tests. AIChE Annual Conference. Paper 435e. http://www.ou.edu/class/che-design/unpub-papers/grosserr-consistency.pdf

Bakshi, B. R. (1998). Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE Journal*, **44**(7), 1596-1610.

Barnet, V. & T. Lewis (1994). *Outliers in Statistical Data*, John Wiley & Sons.

Bishop, C. (1998). Bayesian PCA. *Neural Information Processing Systems (NIPS)*, **11**, 382-388.

Box, G. E. P. (1954). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems – I. *Ann. Math. Stat.*, **25**, 290-302.

Box, G. & D. Cox (1964). An Analysis of Transformations. *In Journal of the Royal Statistics Society*, **26**, 211-252.

Burnham, A. J., R. Viveros, and J. F. MacGregor (1996). Frameworks for latent variable multivariate regression. *J. Chemometr.*, **10**, 31-45.

Charpentier, V., L. J. Chang, G. M. Schwenzer, and M. C. Bardin (1991). An On-line Data Reconciliation System for Crude and Vacuum Units. *NPRA Computer Conference*, CC-91-139, Houston, Tex.

Chen, J. and R. J. Patton (1999). *Robust Model-Based Fault Diagnosis for Dynamic Systems.* Boston, MA: Kluwer.

Choudhury, M. A. A. S., S. L. Shah, and N. F. Thornhill (2004). Diganosis of Poor Control

Loop Performance using Higher Order Statistics. accepted for publication in *Automatica*, June.

Çinar, A. & C. Ündey (2002). Statistical Monitoring of Multiphase Batch Processes. *IEEE Control Systems Magazine*, **Oct**, 40-52.

Crowe, C. M., Y. A. Garcia Campos, and A. Hrymak (1983). Reconciliation of Process Flow Rates by Matrix Projection. I. The Linear Case. *AIChE Journal*, **29**, 881~888.

Crowe, C. M. (1988). Recursive identification of gross errors in linear data reconciliation. *AIChE J.*, **34**, 541-550.

Crowe, C. M. (1989). Test of Maximum Power for Detection of Gross Errors in Process Constraints. *AIChE J.*, **35**, 869~872.

DeBoer, P. & V. Feltkamp (2000). Robust multivariate outlier detection. *Technical Report 2*, Statistics Netherlands, Dept. of Statistical Methods. http://neon.vb.cbs.nl/rsm/P_su2000.htm.

Dehon, C., P. Filzmoser and C. Croux (2000). Robust methods for canonical correlation analysis. In *H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen, M. Schrader, editors, Data Analysis, Classification, and Related Methods,* 321-326, Springer-Verlag, Berlin.

Dunia, R., S. J. Qin, T. F. Edgar, and T. J. McAvoy (1996). Identification of Faulty Sensors Using Principal Component Analysis. *AIChE J.*, **42**(10), 2797-2812.

Everson, R. & S. Roberts (2000). Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Trans Signal Processing*, **48** (7), 2083-2091.

Frank, P. M. (1990). Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy—A survey and some new results. *Automatica*, **26**, 459–474.

Filzmoser, P. (2001). Robust principal component regression. In *S. Aivazian, Y. Kharin, and H. Rieder, editors, Proceedings of the Sixth International Conference on Computer Data Analysis and Modeling,* **1**, 132-137, Minsk, Belarus.

Geladi, P. & B. R. Kowalski (1986a). Partial least-squares regression: A tutorial. *Analytica Chemica Acta,* **185**, 1-17.

Geladi, P. & B. R. Kowalski (1986b). An example of 2-block predictive partial least-squares regression with simulated data. *Analytica Chemica Acta,* **185**, 19-32.

Gertler, J. J. (1988). Survey of model-based failure detection and isolation in complex plants. *IEEE Control Syst. Mag.*, **8**, Dec., 3–11.

Golub, G. H. & C. F. Van Loan (1980). An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, **17** (6), 883-893.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples, *Technometrics*, **11**,1-21.

Hair, J., R. Anderson, R. Tatham and W. Black (1998). *Multivariate Data Analysis*, Prentice Hall International.

Hand, D. J. (1981). *Discrimination and Classification*, Wiley.

Hand, D. J. (1982). *Kernel Discriminant Analysis*, Wiley.

Hastie, T. & W. Stuetzle (1989). Principal curves. *Journal of American Statistical Association*, **84**, 502-516.

Hawkins, D. M. (1980). *Identification of Outliers*, Chapman and Hall, London.

Hawkins, D. M. (1974). The detection of errors in multivariate data using principal components. *Journal of American Statistics Association*, **69**, 340-344.

Hoerl, R. W. (June 1998). Six Sigma and the future of the quality profession. *Quality Progress*. **31**(6), 35-42.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, **24**, 417-441.

Hunter, J. S. (1986). The exponentially weighted moving average. *Journal of Quality Technology* **18**(4), 203-210.

Ibrahim, K. A., and M. T. Tham (1995). Towards active statistical process control. *Proceedings of the American Control Conf.*, **3**, 2234-2238.

Jackson, J. E. (1991). *A user's guide to principal components*. Wiley Series on Probability and Statistics, John Wiley and Sons.

Jansson, Åsa, J. Röttorp and M. Rahmberg (2002). Development of a Software Sensor for Phosphorous in Municipal Wastewater. *Journal of Chemometrics*, **16**, 542-547.

Jia, F., E. B. Martin and A. J. Morris (2000). Non-Linear Principal Components Analysis with application to Process Fault detection. *Int. Journal of System Science*, **31**, 1473 - 1487.

Jiang, Q., M. Sanchez, and M. J. Bagajewicz (1999). On the Performance of Principal Component Analysis in Multiple Gross Error Identification, Ind. & Eng. Chem. Research, **38** (5), 2005-2012.

Jiang, T., B. Chen, X He, and P. Stuart (2003). Application of steady-state detection method based on wavelet transform. *Computers & Chemical Engineering*, **27** (4), 569-578.

Johnson, R. A. and D. W. Wichern (1998). *Applied Multivariate Statistical Analysis*. 4th. ed., Prentice-Hall.

Jordache, C., R. S. H. Mah, and A. C. Tamhane (1985). Performance Studies of the Measurement Test for Detecting Gross Errors in Process Data.AIChE Journal, **31**, 1187-1201.

Jordache, C., and B. Tilton (1999). Gross Error Detection by Serial Elimination: Principal Component Measurement Test versus Univariate Measurement Test. presented at the *AIChE Spring National Meeting*, Houston, Tex., March.

Kleiner, B. and J. A. Hartigan (1981). Representing points in many dimensions by trees and castles (with discussion), *Journal of American Statistics Association*, **76**, 260-276.

Kourti, T., P. Nomikos, and J.F. MacGregor (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control.* **5**(4), 277-284.

Kourti, T. (2002). Process Analysis and Abnormal Situation Detection: from theory to practice. *IEEE Control Systems Magazine*, **Oct.**, 10-25.

Kramer, M. A. (1991). Nonlinear Principal Component Analysis Using Autoassociative Neural Networks, *AIChE J.*, **37**, 23-243.

Kramer, M. A., and R. S. H. Mah (1994). Model-Based Monitoring, in *Proc. Second Int. Conf. On Foundations of Computer Aided Process Operations*, D. Rippin, J. Hale, J. Davis, eds.CACHE.

Kresta, J. V., J.F. MacGregor and T.E. Marlin (1991). Multivariate Statistical Monitoring of Process Operating Performance, *Canadian Journal of Chemical Engineering*, **69**, 35-47.

Ku, W., R. H. Storer, and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, **30**(1), 179-196.

Kuehan, D. R., and H. Davidson (1961). Computer Control. II. Mathematics of Control. *Chem. Eng. Progress*, **57**, 44-47.

Lakshminarayanan, S. (1997). Process Characterization and Control using Multivariate Statistical Techniques. PhD thesis. University of Alberta. Canada.

Larimore, W. E., S. Mahmood and R.K. Mehra (1984).Multivariable Adaptive Model Algorithmic Control. *Proceedings of the Conference on Decision and Control*, **2**, 675-680.

Larimore, W. E. (1990). Canonical Variate Analysis in Identification, Filtering and Adaptive Control. *Proceedings of the 29th Conference on Decision and control*, Hawai, USA, pp. 596-604.

Li, G., Chen Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Amer. Statist. Assoc.*, **80**(391), 759-766.

Li, W., H. Yue, S. Valle, S. J. Qin (2000). Recursive PCA for adaptive process monitoring. *Journal of Process Control*, **10**, 471-486.

Li, W., S. Joe Qin (2001). Consistent dynamic PCA via errors-in-variables subspace identification. *Journal of Process Control*, **11**(6), 661-678.

Liu, H., S. Shah, and W. Jiang (2003). Outlier detection and data cleaning, Submitted to Computers and Chemical Engineering.

Liu, X., G. Cheng, and J. Wu (1994). Managing the noisy glaucomatous test data by selforganizing maps. *Proceedings of IEEE International Conference on Neural Networks*, 649-652, Orlando, Florida.

Liu, X., G. Cheng, and J. Wu (1998). Managing the noisy glaucomatous test data by self organizing maps, *Proceedings of IEEE International Conference on Neural Networks*, 649-652.

Liu, X., G. Cheng,., and J. Wu (2001), Analyzing outliers cautiously, *IEEE Transactions on Knowledge and Data Engineering*, to appear.

Liu, X., G. Cheng, and J. Wu (2002). Analyzing outliers cautiously. *IEEE Transactions on Knowledge and Data Engineering*, **14**(2), March-April, 432 –437.

Lucas, J. M. and M.S. Saccucci (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technomitrics* **32** (1), 1-12.

Luo, R. F., Misra M. and Himmelblau D. M. (1999). Sensor Fault Detection via Multiscale Aanalysis and Dynamic PCA. *Ind. Eng. Chem. Res.*, **38**, 1489-1495.

MacGregor, J. F., T. E. Marlin, J. V. Kresta and B. Skagerberg (1991)., Multivariate Statistical Methods in Process Analysis and Control, *Proc. of Fourth Intl. Conf. On Chemical Process Control*, 665-672.

MacGregor, J. F., C. Jaeckle, C. Kiparissides, and M. Koutoudi (1994a). Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.*, **40**(5), 826-838.

MacGregor, J. F. (1994b). Statistical Process Control of Multivariate Processes, *IFAC ADCHEM*, Kyoto, Japan.

Madron, F. (1985). A new approach to the identiÞcation of gross errors in chemical engineering measurements. *Chem. Eng. Sci.*, **40**, 1855-1860.

Madron, F. (1992). *Process Plant Performance: Measurement and Data Porcessing for Optimization and Retrofits*. Chichester, West Sussex, England, Ellis Horwood Limited Co.

Mah, R. S. H., Stanley, G.M. and D. M. Downing (1976). Reconciliation and RectiÞcation of Process Flow and Inventory Data. *Ind. Eng. Chem. Process Des. Dev.*, **15**, 175-183.

Mah, R. S. H. and A. C. Tamhane (1982). Detection of Gross Errors in Process Data. *AIChE J.*, **28**, 828-830.

Mah, R. S. H. (1990). *Chemical Process Structures and Information Flows*, Butterworths, Boston.

Malinowski, E. R. (1991). *Factor Analysis in Chemistry*. Wiley-Interscience, New York.

Meng, X., E. B. Martin and A. J. Morris (2000). A Comparative Study of Bi-linear and Tri-linear Approaches for the Monitoring of an Industrial Batch Process. ESCAPE-10: European Symposium on Computer Aided Process Engineering, Florence, Italy, 7-10 May.

Miller, P., R. E. Swanson and C. F. Heckler (1993). Contributions plots: the missing link in multivariate quality control. *37th Annual Fall Conf. ASQC* (Rochester, N.Y).

Minka, T. P. (2000). Automatic Choice of Dimensionality for PCA. *Neural Information Processing Systems (NIPS)*, Denver, CO, USA, 598-604.

Misra, M., S. J. Qin, H. Yue and C. Ling (2002). Multivariate process monitoring and fault identification using multi-scale PCA. *Computers and Chemical Engineering*, **26**, 1281–1293

Montague, G. A., H. G. Hiden, and G. Kornfeld (1998). Multivariate statistical monitoring procedures for fermentation supervision: an industrial case study. *Proc. Conf. Comp. App. In Biotechnology*, Osaka, Japan.

Montgomery, D. C. (1996). Introduction to Statistical Quality Control, 3th ed. New York: Wiley.

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y. (2000a). Statistical software ``VASPCA" for variable selection in principal component analysis, In: W. Jansen and J.G. Bethlehem (eds.) *COMPSTAT2000 Proceedings in Computational Statistics* (*Short Communications*), 73-74. http://mo161.soci.ous.ac.jp/vaspca/indexE.html

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

Munoz, A. and J. Muruzábal (1998), Self-organizing maps for outlier detection, *Neurocomputing*, 18:33-60.

Narasimhan, S. and R. S. H. Mah (1987). Generalized likelihood ratio method for gross error detection. *A.I.Ch.E.J.*, **33**, 1514.

Narasimhan, S. and R. S. H. Mah (1988). Generalized likelihood ratios for gross error identiÞcation in dynamic processes, *A.I.Ch.E. J.*, **33**, 1514.

Narasimhan, S. and C. Jordache (2000). *Data Reconciliation and Gross Error Detection-An Intelligent Use of Process Dat.* Gulf Publishing Company.

Narasimhan, S. and S. L. Shah (2004). Model Identification and Error Covariance Matrix Estimation From Noisy Data Using PCA, *Proc. of IFAC Adchem conference, Hong Kong, Jan 2004, 567-572.*

Nelson, Philip R. C., P. A. Taylor and J. F. MacGregor (1996). Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations. *Chemometrics and Intelligent Laboratory Systems*, **35**(1), 45-65.

Nomikos, P., and J. F. MacGregor (1994). Monitoring Batch Processes Using Multiway Principal Component Analysis, *AIChE J.*, **40** (8), 1361.

Nounou, M. N., B. R. Bakshi, P. K. Goel, and X. Shen (2002). Bayesian principal component analysis. *Journal of Chemometrics*, **16** (11), 576-595.

Orlando, Florida, Munoz, A. and J. Muruzábal (1994). Self-organizing maps for outlier detection. *Neurocomputing*, **18**, 33-60.

Oxby, P. W. and S. L. Shah (2000). A Critique of the Use of PCA for Fault Detection and Diagnosis. *Research report, U. of Alberta, 2002.*

Page, E. S. (1954). Continuous inspection schemes. *Biametrika*, **41**, 100-115.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *The London, Edinburg and Dublin Philosophical Magazine and Journal of Science, Sixth Series,* **2**, 559-572.

Phatak, A. and S. D. Jong (1997). The geometry of partial least squares. *Journal of Chemometrics*, **11**, 311–338.

Rajan, J. J. and P. J. W. Rayner (1997). Model order selection for the singular value decomposition and the discrete Karhunen-Loéve transform using a Bayesian approach. *IEE Proceedings - Vision, Image and Signal Processing*, **144** (2), 166-123.

Reilly, P. M. and Carpani, R. E. (1963). Application of statistical theory of adjustment to material balances. *13th Can. Chem. Eng. Conf.*, Montreal, Quebec.

Ripps, D. L. (1965). Adjustment of Experimental Data. *Chem. Eng. Progress Symp.*, Series **61**, 8~13.

Roberts, S.W. (1959). Control chart testes based on geometric moving averages. *Technomitrics* 1, 239-250.

Rollins, D. K., Y. Cheng and S. Devanathan (1996). Intelligent Selection of Hypothesis tests to Enhance Gross Error Identification. *Comp. and Chem. Eng.*, **20**(5), 517-530.

Rollins, D. K. and J. F. Davis (1992). Unbiased estimation of gross errors in process measurements. *A.I.Ch.E. J.*, **38**, 563.

Rosenberg, J., R. S. H. Mah, and C. Jordache (1987). Evaluation of Schemes for Detecting and Identifying Gross Errors in Process Data. *Ind. & Eng. Chem. Proc. Des. Dev.*, **26**, 555-564.

Rouseeuw, P. J. and K. V. Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**(3), 212-223.

Rumantir, G. W. (1995). A Hybrid Statistical and Feedforward Network Model for Forecasting with a Limited Amount of Data: Average Monthly Water Demand Time-series, *MSc. Thesis*, Department of Computer Science. Royal Melbourne Institute of Technology (RMIT University).

Salvador, G. M., T. Kourti, J. F. MacGregor, G. Arthur, and G. Murphy (2003). Troubleshooting of an industrial Batch Process Using Multivariate Methods. *Ind. Eng. Chem. Res.*, **42**, 3592-3601.

Sánchez, M., and J. Romagnoli (1996). Use of Orthogonal Transformations in Data Classification-Reconciliation. *Computers chem. Engng.*, **20**, 483-493.

Sánchez, M., J. Romagnoli, Q. Jiang, and M. Bagajewicz (1999). Simultaneous Estimation of biases and Leaks in Process Plants. *Computers & Chem. Eng.*, **23** (7), 841-858.

Sánchez, M. and J. Romagnoli (2000). *Data Processing and Reconciliation for Chemical Process Operations*. Academic Press.

Schaper, C.D., W.E Larimore, D.E. Seborg and D.A. Mellichamp (1994), Identification of Chemical Processes using Canonical Variate Analysis. *Computers and Chemical Engineering*, **18**, 55-69.

Schölkopf, B., Smola A. J., and K. R. Müller (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*. **10**, 1299-1319.

Serth, R. and W. Heenan (1986). Gross error detection and data reconciliation in steam metering systems. *A.I.Ch.E.J.*, **32**, 733.

Shao, R., F. Jia, E. B. Martin and A. J. Morris (1999). Wavelets and Non-linear Principal Components Analysis for Process Monitoring. *Control Engineering Practice*, **7**, 865-879.

Shewart, W.A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand, Princeton, N.J.

Shi, D. and F. Tsung (2003). Modeling and Diagnosis of Feedback-Controlled Processes Using Dynamic PCA and Neural Networks. *International Journal of Production Research*, **41**, 365-380.

Sidak, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J. Amer. Statist. Assoc.*, **62**, 626.

Skočaj, D., H. Bischof, and A. Leonardis (2002). Robust PCA Algorithm for Building Representations from Panoramic Images. *Proceeding of ECCV*, May, 28-31, Copenhagen, Danemark.

Sumpter, N., R. D. Boyle, and R. D. Tillet (1997). Modelling collective animal behaviour using extended point distribution models. In A. F. Clark, editor, *British Machine Vision Conference*, 1, 242–251.

Swartz, C. L. E. (1989). Data Reconciliation for Generalized Flowsheet Applications. *American Chemical Society National Meeting*, Dallas, Tex.

Tamhane, A. C. and R. S. H. Mah (1985). Data reconciliation and gross error detection in chemical process networks. *Technometrics*, **27**(4), 409.

Tamhane, A. C., C. Jordache, and R. S. H. Mah (1988). A Bayesian Approach to Gross Error Detection in Chemical Process Data. Part I: Model Development. *Chemometrics and Intel. Lab. Sys.*, **4**, 33-45.

Tanaka, Y. and Mori, Y. (1997). Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematics and Management Sciences*, **17**(1&2), 61-89.

Tong, H. and C. M. Crowe (1995). Detection of Gross Errors in Data Reconciliation by Principal Component Analysis. *AIChE J.*, **41**(7), 1712-1722.

Tong, H. and C. Crowe (1996). Detecting Persistent Gross Errors by Sequential Analysis of Principal Components. *Comp. Chem. Eng.*, **20**, 733-738.

Tong, H. and D. Bluck (1998). An Industrial Application of Principal Component Test to Fault Detection and Identification. *IFAC Workshop on On-Line-Fault Detection and Supervision in the Chemical Process Industries*, Solaize (Lyon), France.

Tipping, M. E. and C. M. Bishop (1999a). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, **Series B-61** (3), 611-622.

Tipping, M. E. and C. M. Bishop (1999b). Mixtures of probabilistic principal component analyzers. *Neural Computation*, **11**(2), 443-482. http://research.microsoft.com/users/mtipping/pages/publications_abs.htm

Tsung, F. (2000). Statistical Monitoring and Diagnosis of Automatic Controlled Processes Using Dynamic PCA. *International Journal of Production Research*, **38**, 625-637.

Vaclavek, V. (1968). Studies on System Engineering. I. On the Application of the Calculus of Observations in Calculations of Chemical Engineering Balances. *Coll. Czech. Chem. Commun*, **34**, 3653.

Vaclavek, V., and M. Loucka (1976). Selection of Measurements Necessary to Achieve Multicomponent Mass Balances in Chemical Plant. *Chem. Eng. Sci.*, **31**, 1199-1205.

Valle, S., W. Li, and S. J. Qin (1999). Selection of the Number of Principal Components: the Variance of the Reconstruction Error Criterion with Comparison to Other Methods. *Ind. Eng. Chem. Res.*, **38**, 4389-4401.

Venkatasubramanian, V., R. Rengaswamy, K. Yin, and S. N. Kavuri (2003). A review of process fault detection and diagnosis, Part I: Quantitative model-based methods. *Computers and Chemical Engineering*. **27**, 293-311.

Wangen, L. E. and B. R. Kowalski (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, **3**, 3-20.

Wansbeek, T. and E. Meijer (2000). Measurement Error and Latent Variables in Econometrics. *Elsevier Science B. V.*, The Netherlands.

Wentzell, P. D., D. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski (1997). *J. of Chemometrics*, **11**, 339-366.

Westerhuis, J. A., P. M. J. Coenegracht, and C. F. Lerk (1997). Multivariate modelling of the tablet manufacturing process with wet granulation for tablet optimization and in-process control. *International Journal of Pharmaceutics*, **156**, 109-117.

Westerhuis, J. A., T. Kourti and J. F. MacGregor (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, **12**, 301-321.

Wise, B. M., N. L. Ricker, D. F. Veltkamp and B. R. Kowalski (1990). A theoretical Basis for the Use of Principal Component Models for Monitoring Multivariate Processes. *Process Control and Quality*, **1**, 41-51.

Wise, B. M., and N. B. Gallagher (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, **6**(6), 329-348.

Willsky, A. S., and H. L. Jones (1974). A Generalized Likelihood Ratio Approach to State Estimation in Linear Systems Subject to Abrupt Changes. *proc. IEEE Conf. Decision and Control*, 846.

Wold, H. (1966). *Multivariate Analysis*. Academic, New York.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technomitrics*, **20**, 397-405.

Wold, S. (1994). Exponentially weighted moving principal component analysis and projection to latent structures. *Chemometrics and Intelligent Laboratory Systems*, **23**, 149-161.

Woodward, R. H. and P. L. Goldsmith (1964). *Cumulative Sum Techniques*. Published for Imperial Chemical Industries, Oliver Boyd, London, England.

Yoon, S. and J. F. MacGregor (2001). Fault diagnosis with multivariate statistical models, Part I: Using steady state fault signature. *Journal of Process Control*, **11**, 387-400.

Yu Qian, H. Cheng, X. Li, and Y. Jiang (2002).. Dynamic Process Modelling using a PCA-based Output Integrated Recurrent Neural Network. *The Canadian Journal of Chemical Engineering*, **80**, 1-6.

Yue, H. and S. J. Qin (2001). Reconstruction based fault identification using a combined index. *I&EC Research*, **40**, 4403-4414.

Zhang, H. (2000). Statistical Process Monitoring using PCA and PLS. MSc. thesis. University of Alberta. Canada.

Appendix A

Given the $n\times1$ measurement vector $\mathbf{y}$ of the true values of variables $\mathbf{x}$ and the measurement error $\mathbf{e}$, i.e., $\mathbf{y}=\mathbf{x}+\mathbf{e}$, and $\mathbf{x}$ conforms strictly to the model $\mathbf{A}^*$ as $\mathbf{A}^*\mathbf{x}=\mathbf{0}$, if $\mathbf{e}\sim N(\mathbf{0},\Sigma_e)$ and the error covariance $\Sigma_e=\sigma^2\mathbf{I}$, then the following statement is true:

If PCA is applied to the zero-mean data matrix $\mathbf{YY}^T/(N-1)$, where $\mathbf{Y}=\mathbf{X}+\mathbf{E}$ is the N-sample realization of $\mathbf{y}=\mathbf{x}+\mathbf{e}$, then the transform of loading vectors $\mathbf{B}=(v_{k+1}, v_{k+2}, \ldots, v_n)$ in equation 2.3 corresponding to the $m=n-k$ small eigenvalues will asymptotically be the similarity transformation (or full-rank linear transformation) of $\mathbf{A}^*$ when $N\to\infty$.

Proof

(1) For $\sigma^2=0$, then $\mathbf{Y}=\mathbf{X}$, the rank of $\mathbf{YY}^T$ is same as the rank of $\mathbf{XX}^T$. If we apply PCA to the sample covariance matrix $\mathbf{YY}^T/(N-1)$, the transpose of $\mathbf{B}$, the last m orthonormal eigenvectors corresponding to the zero eigenvalues represent a basis for the residual space (RS), which is orthogonal to the data vectors $\mathbf{Y}$. If we denote this basis for RS by a m dimensional linear model $\mathbf{B}^T$, then we have:

$$\mathbf{B}^T\mathbf{Y}=\mathbf{B}^T\mathbf{X}=\mathbf{0}$$

$\mathbf{B}^T$ and $\mathbf{A}^*$ hold the relationship as:

$$\mathbf{MB}^T=\mathbf{A}^*$$

where $\mathbf{M}=\mathbf{A}^*\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}=\mathbf{A}^*\mathbf{B}$ is a full rank $m\times m$ matrix.

(2) For $\sigma^2>0$, then $\mathbf{Y}=\mathbf{X}+\mathbf{E}$, where $\mathbf{YY}^T$ has full rank.

When $N\to\infty$, $\mathbf{YY}^T/(N-1)=\mathbf{XX}^T/(N-1)+\Sigma_e$, applying PCA we have

$$\mathbf{VSV}^T=\mathbf{XX}^T/(N-1)+\Sigma_e \Rightarrow \mathbf{S}=\left(\frac{\mathbf{X}^T}{\sqrt{N-1}}\mathbf{V}\right)^T\frac{\mathbf{X}^T}{\sqrt{N-1}}\mathbf{V}+\mathbf{V}^T\Sigma_e\mathbf{V}$$

$$\Rightarrow\left(\mathbf{x}^T\mathbf{V}\right)^T\mathbf{x}^T\mathbf{V}=\mathbf{S}-\mathbf{V}^T\Sigma_e\mathbf{V} \qquad \ldots\ldots\ldots *$$

$$\Rightarrow\left(\mathbf{x}^T\mathbf{V}\right)^T\mathbf{x}^T\mathbf{V}=\mathbf{S}-\sigma^2\mathbf{I}$$

Here the quadratic form $\left(\frac{\mathbf{X}^T}{\sqrt{N-1}}\mathbf{V}\right)^T\frac{\mathbf{X}^T}{\sqrt{N-1}}\mathbf{V}$ and $\left(\mathbf{x}^T\mathbf{V}\right)^T\mathbf{x}^T\mathbf{V}$ are asymptotically equivalent.

$\mathbf{V}$ is the matrix of loading vectors and $\mathbf{S}$ is the diagonal eigenvalue matrix with descending order. Because the rank of $\mathbf{XX}^T$ deducts by m order, if the signal-to-noise ratio SNR is significant in $\mathbf{X}$, then, according to equation *, $\mathbf{S}$ takes the form of as below:

$$\mathbf{S} = \begin{bmatrix} \overbrace{\begin{matrix} s_1 & & \\ & s_2 & \\ & & \ddots \end{matrix}}^{k} & \mathbf{0} \\ \mathbf{0} & \overbrace{\begin{matrix} s_k & & \\ & \sigma^2 & \\ & & \ddots \end{matrix}}^{m} \\ & \qquad\qquad \sigma^2 \end{bmatrix} \quad \text{..... **}$$

$n \times n$ diagonal

where $s_1 > s_2 \cdots > s_k > \sigma^2$.

Thus, for the first k columns of $\mathbf{V}$, $\mathbf{P} = (v_1, v_2, \ldots, v_k)$, $\left(\mathbf{x}^T v_i\right)^T \mathbf{x}^T v_i > 0$, where i=1~k; for the last m=n-k columns of $\mathbf{V}$, $\mathbf{B} = (v_{k+1}, v_{k+2}, \ldots, v_n)$, $\left(\mathbf{x}^T v_i\right)^T \mathbf{x}^T v_i = 0$, where i=k+1~n. Collectively rewriting this result gives: $\mathbf{B}^T\mathbf{x} = \mathbf{0}$. Comparing to the given model $\mathbf{A}^*\mathbf{x} = \mathbf{0}$ we obtain $\mathbf{MB}^T = \mathbf{A}^*$, where $\mathbf{M}$ is the same as in (1).

From (1) and (2) we conclude that the row orthonomal matrix $\mathbf{B}^T = \mathbf{V}_{1+k\sim n}^T$ is a **linear transformation** of $\mathbf{A}^*$, where $\mathbf{B}$ is the last m=n-k eigenvectors of $\mathbf{YY}^T/(N-1)$, corresponding to all the small eigenvalues. $\mathbf{Y}$ is a N-sample measurement data set.