STUDIES IN DATA RECONCILIATION USING PRINCIPAL COMPONENT ANALYSIS

By

Hongwei Tong, B.Eng., M.S.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of Requirements

for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (1995) McMASTER UNIVERSITY

(Chemical Engineering) Hamilton, Ontario, Canada

TITLE:	Studies in Data Reconciliation Using Principal Component Analysis
AUTHOR:	Hongwei Tong, B.Eng.; M.S. (Tsinghua University)
SUPERVISOR:	Professor Cameron M. Crowe
NUMBER OF PAGES:	xii, 149

Studies in Data Reconciliation Using Principal Component Analysis

ABSTRACT

Measurements such as flow rates from a chemical process are inherently inaccurate. They are contaminated by random errors and possibly gross errors such as process disturbances, leaks, departure from steady state, and biased instrumentation. These measurements violate conservation laws and other process constraints. The goal of data reconciliation is to resolve the contradictions between the measurements and their constraints, and to process contaminated data into consistent information. Data reconciliation aims at estimating the true values of measured variables, detecting gross errors, and solving for unmeasured variables.

This thesis presents a modification of a model of bilinear data reconciliation which is capable of handling any measurement covariance structure, followed by a construction of principal component tests which are sharper in detecting and have a substantially greater power in correctly identifying gross errors than the currently used statistical tests in data reconciliation. Sequential Analysis is combined with Principal Component Analysis to provide a procedure for detecting persistent gross errors.

The concept of zero accumulation is used to determine the applicability of the established linear/bilinear data reconciliation model and algorithms. A two stage algorithm is presented to detect zero accumulation in the presence of gross errors.

An interesting finding is that the univariate and the maximum power tests can be quite poor in detecting gross errors and can lead to confounding in their identification.

ACKNOWLEDGMENTS

I wish to express sincere appreciation to my supervisor, Dr. Cameron M. Crowe, for his guidance, patience, encouragement and many valuable discussions throughout this investigation.

I also wish to thank:

Drs. John F. MacGregor, Andrew N. Hrymak and James P. Reilly for their valuable input and encouragement as the members of my supervisory committee.

Dr. Thomas E. Marlin for his input to my research proposal, Dr. Theodora Kourti and Ms. Meiying Hou for their help with PCA, Mr. Oliver Schraa for his help with numerical applications and many challenging discussions, and Dr. Fraser Forbes for his encouragement and discussions on model mismatch.

I wish to express my gratitude to Prof. Xianshun Zeng of Tsinghua University for his help with the generalized hypergeometric function.

Sincere appreciation is extended to Ms. Sara Gallo-O'Toole, Ms. Barbara Owen and Ms. Sharon Ciraolo for their support and assistance throughout my studies at McMaster.

To my parents, I thank you for your constant interest and encouragement.

To my wife, Fei Zhao, I offer sincere gratitude and appreciation for her unselfish support, patience and understanding during the completion of this work.

To my little daughter, Theresa, I thank you for all the brightness and happiness that you brought to me during the years of my study.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

Measurements such as flow rates from a chemical process are inherently inaccurate. They are contaminated by random errors and possibly gross errors. Such errors may arise from, for instance, process disturbances, leaks, malfunctioning or miscalibrated instrumentation, and departure from steady state. These measurements violate conservation laws and other process constraints. The aim of data reconciliation is to resolve the contradictions between the measurements and their constraints, and to process contaminated data into consistent information.

In this thesis we investigate how to reconcile data contaminated by random errors, how to detect and identify non-random (gross) errors including persistent gross errors in the data, how to improve sharpness and reduce confounding in statistical tests, and how to determine periods of steady state so that steady state data reconciliation can be performed.

A general review of data reconciliation can be found in Crowe (1994). Other excellent sources are Crowe *et al.* (1983), Crowe (1986), Mah (1987, 1990), Madron (1992), Ragot *et al.* (1992), and Aldrich and Van Deventer (1994). A good introduction for engineers can be found in Lawrence (1989).

This chapter gives an overview and addresses some fundamental problems of data reconciliation.

1.1. Objectives of Data Reconciliation

Data reconciliation aims at estimating the true values of measured variables, detecting gross errors, and solving for unmeasured variables. They are respectively an optimization problem, a statistical hypothesis testing problem, and an equation-solving problem with the number of unknowns being less than the number of the equations.

We should emphasize that without performing statistical tests and computing unmeasured quantities, data reconciliation may be invalid and infeasible. This problem has been overlooked in some commercial software packages and publications.

1.2. Applications of Data Reconciliation

Data reconciliation is useful in many areas such as overall mass accounting, process monitoring and analysis, model identification, process control and optimization, production planning, and detection of faulty instrumentation and process leaks. In fact, for improving confidence in measurements, data reconciliation can be applied to any problem where measurements and process constraints are both available. Recent industrial applications include data reconciliation of measurements in a refinery (Albers, 1994), an ethylene plant (Sanchez *et al.*, 1992), a pulp production process (Markos and Barto, 1991), and a chemical extraction plant (Holly *et al.* 1989). Islam *et al.* (1994) applied data reconciliation to an industrial pyrolysis reactor. Van der Heijden *et al.*, (1994a, 1994b) used data reconciliation to optimally adjust conversion rates in a biochemical reactor. Bossen *et al.* (1994) employed data reconciliation to obtain heat transfer coefficients in heat exchangers and relative activity for the catalyst in an ammonia reactor, in order to develop short-cut models for optimizing the ammonia synthesis process. Simultaneous data reconciliation and parameter estimation was discussed in Pages *et al.* (1994) and Kim *et al.* (1990).

1.3. Variables for Data Reconciliation

Some basic variables such as total flow rates, temperature, and concentrations can be directly measured. Others are either difficult to measure (e.g., catalyst activity) or too expensive to measure. Furthermore, a model of steady state material and energy balances in terms of those basic variables is much easier to build and more accurate than one in terms of other variables. Such a steady state balancing model is usually available in commercial simulators. Therefore a good data reconciliation program can be integrated into those general purpose simulators.

We may classify directly measured variables as primary variables, and the others as secondary variables. It is more efficient to reconcile the primary variables than to reconcile the secondary variables. Once the measurements are reconciled, the secondary variables can be computed. Hence, most practical problems can be reduced to reconciliation of the primary variables.

A frame work of steady state data reconciliation is given in Chapter 2.

1.4. Knowledge Needed for Data Reconciliation

To achieve the objectives of data reconciliation, certain process knowledge is required.

Measurements and a process model are the fundamental prerequisites of data reconciliation. With the increase in on-site data collection, the problems of access to data are rapidly diminishing. The problem left is data inconsistency (Lawrence, 1989). The inconsistency can only be resolved against a reliable process model. Such a model usually consists of steady state material and energy balances and equilibrium relationships. Models

of dynamic processes often involve parameters that are uncertain. Such uncertainties complicate reconciliation of dynamic processes.

If we assume that a model is accurate, we can reconcile data against the model. This is a data reconciliation problem. If we assume that data are accurate, we can estimate the model parameters. This is a parameter estimation problem. A steady state model consisting of material and energy balances is accurate, indeed exact, in general. However, model mismatch may occur if flows or reactions are overlooked.

Some researchers performed dynamic data reconciliation using general nonlinear process constraints (Liebman *et al.*, 1992) and Kalman filter (Almasy, 1990). Unless the parameters in the nonlinear constraints and Kalman filter are much more accurate than the measurements, model mismatch is inevitable and reconciliation is unreliable.

Another prerequisite of data reconciliation is a measurement model. Measurement errors are conveniently assumed to follow a normal distribution. Some researchers investigated problems when the errors do not follow a normal distribution (Tjoa and Biegler, 1991; Kao *et al.*, 1990). Recently, Crowe (1995) proposed a theory on linear data reconciliation with the minimum incorporation of prior knowledge about probability distributions of the data by maximizing the information entropy. In his formulation, reconciliation is achievable without prior knowledge of the variance-covariance of the data.

The last yet not the least important prerequisite of data reconciliation is measurement redundancy (Crowe, 1989a; Stanley and Mah, 1981). A measured quantity is redundant if and only if it could be uniquely calculated from the other measurements, had it not been measured. Non-redundant measurements have no effect on data reconciliation

and are only useful in calculating unmeasured variables. Without redundancy, there would be no data reconciliation, because there would be no degrees of freedom for any adjustments to the measured quantities. Sometimes redundancy is available only in a part of a process, so that data reconciliation can be done only for that part of the process.

A concept that closely relates to redundancy is observability (Crowe, 1989a; Stanley and Mah, 1981; Maquin *et al.*, 1989). An unmeasured quantity is observable if and only if it can be deduced from a consistent set of measurements. Observability is not a prerequisite for data reconciliation.

1.5. Gross Error Detection and Identification

Two important objectives of data reconciliation are to optimally adjust measurements and detect gross errors. The optimal adjustments to the measurements can be made only when the measurements in gross error have been removed or corrected. Practically, a preliminary reconciliation is performed with all of the measurements. Hypothesis testing is then carried out to investigate if there is any evidence of gross errors[1]. When a gross error is found, the corresponding measurement is removed or corrected, and another reconciliation is performed. This procedure is repeated until no evidence of gross errors is found. The unmeasured quantities are then solved for if they are observable, based on the reconciled measurements. It is clear that the hypothesis testing for detecting gross errors also plays a role in validating the adjustments to the measurements.

[1] Certain statistical testing can be done without the preliminary reconciliation, others cannot. As seen from the following chapters, hypothesis testing and data reconciliation share many calculations so that there is no significant saving in computation by not performing the preliminary reconciliation.

There are two fundamental problems associated with data reconciliation: gross error detection and identification. Gross error detection techniques determine whether there is any gross error, while gross error identification techniques find the specific gross errors. These problems have not been well solved, which may have limited the implementation of data reconciliation in industries.

A gross error was described by Van der Heijden *et al.* (1994b) as one that is relatively large compared to the variable's variance. That is not precise because a gross error can be relatively small compared to the variable's variance, as can be seen in some examples in Chapter 5. It would be more appropriate to state that a gross error in a measurement or a constraint violates the process covariance structure. Such a subtle gross error is difficult to detect by currently used statistical methods in data reconciliation.

Quite often, a subtle gross error is not at all the least important error. For instance, it may be crucial to detect a gross error that is small in size but violates a control objective or quality specifications, see Figure 1.1. Once a gross error is detected, it may also be difficult to locate it because the correlation caused by the process topology confound the statistics. Confounding statistics may be misleading and may result in throwing away good data while keeping corrupted ones.

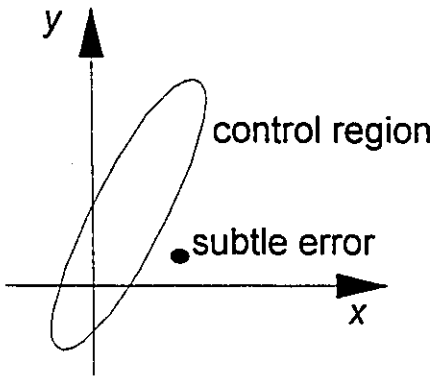


Figure 1.1 A Subtle Gross Error

The principal component tests to be presented in Chapter 4 are helpful in providing sharper detection and better identification of gross errors.

One should be cautious in detecting gross errors solely from statistics. Though using statistics may be the best way to find an error, certain gross errors just could not be detected by statistics. For instance, the flow rate measurements from a splitter shown in Figure 1.2 are significantly biased compared to their true values shown in Figure 1.3. This cannot be detected by any statistic because the material balance is "satisfied". This indicates that the absence of statistical evidence of gross error does not necessarily mean that there is no gross error at all, though it would be unlikely that we encounter such a coincidence in a real process. For simplicity, throughout this thesis we will say that there is no gross error whenever there is no statistical evidence of one.

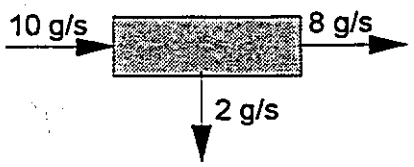


Figure 1.2 Measured Flow Rates

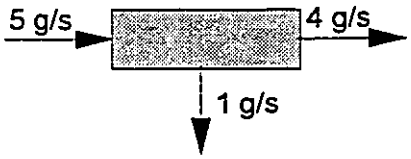


Figure 1.3 True Flow Rates

1.6. About the Thesis

This thesis consists of eight chapters. In Chapter 1, we defined data reconciliation and its objectives, introduced the concepts of model accuracy and mismatch, reviewed the importance and applications of data reconciliation in numerous fields, and discussed the procedures in performing data reconciliation. In Chapter 2, steady state linear and bilinear data reconciliation are reviewed. A generalization of Crowe's model (Crowe, 1986) for data reconciliation is presented to cope with a bilinear problem with arbitrary measurement variance-covariance structure. The ranks for all covariance matrices used in hypothesis testing are given. This is the fundamental information needed to perform principal component tests for detecting gross errors. In Chapter 3, the currently used statistical tests for detecting gross errors are summarized. In Chapter 4, PCA is introduced and the principal component tests are derived to deal with the problems in detecting subtle

gross errors and confounding statistics. Relationships and comparison among the principal component tests and the currently used statistical tests are given. In Chapter 5, a number of examples from chemical, petroleum, and mineral processing industries are shown to illustrate the use of the principal component tests. The differences among different statistical tests are compared. In Chapter 6, Sequential Analysis is reviewed and the sequential principal component tests are presented. The sequential principal component tests are aimed at providing an earlier and sharper detection of gross errors. In Chapter 7, constant and zero accumulations are defined. Sequential Analysis is used to detect periods of zero accumulation in the presence of gross errors when steady state data reconciliation method can be used. In Chapter 8, review of the thesis and recommendations for future research work are addressed.

Appendix A and Appendix B summarize the relationships among the key variables and the variance-covariance matrices in linear and bilinear data reconciliation. These relationships are not well documented in literature yet useful in developing data reconciliation algorithms and establishing the ranks of the variance-covariance matrices. Also shown there are the assumptions used in hypothesis testing for gross errors.

CHAPTER 2

STEADY STATE DATA RECONCILIATION

Data reconciliation is aimed at estimating the true values of measured variables, solving for unmeasured variables if they are observable, and detecting and identifying gross errors. In this chapter, the first two problems are investigated. In particular, a modification of Crowe's model (Crowe, 1986) for data reconciliation is presented to cope with a bilinear problem with an arbitrary measurement variance-covariance structure. The method can be used, but is not limited, to reconcile measurements of total flow rates and concentrations. The ranks for all covariance matrices used in hypothesis testing are given. This information is fundamental in performing principal component tests for detecting and identifying gross errors. Numerical examples are given in Chapter 5.

Measurements violate conservation laws and other process constraints because they are contaminated by errors. The reconciled measurements are consistent in satisfying conservation laws and process constraints. They have smaller confidence regions than those of the original measurements.

It is common that not all streams in a process nor all variables in a stream are measured, yet information about that stream may be highly desirable. If the unmeasured quantities are observable, their estimates can be obtained based on the reconciled measurements.

In chemical engineering, Kuehn and Davidson (1961) were the first to publish an analysis of data reconciliation, presenting the general solution when all flows are measured. A comprehensive review of the history and the state-of-the-art of data reconciliation can be found in Crowe (1994).

2.1. Model of Steady State Data Reconciliation

We now build a model for steady state data reconciliation. Following Crowe *et al.* (1983) and Crowe (1986), we consider a chemical plant in which there are K process units, J streams and up to C components in any stream. The structure of the plant can be expressed in the incidence matrix A with rows corresponding to units and columns to streams. The entry A_{kj} is 1 if stream j enters unit k, -1 if stream j leaves unit k, and 0 otherwise, for $k = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, J$. The balance matrix $\boldsymbol{B}$ consisting of the coefficients in balancing equations can be obtained by replacing each ± 1, 0 in $\boldsymbol{A}$ by the $\pm$ identity and null matrices, respectively. Occasionally matrix $\boldsymbol{B}$ has to be constructed directly from the flowsheet rather than through $\boldsymbol{A}$, such as the Ripps' example (Crowe *et al.*, 1983).

If $\boldsymbol{x}$ is the vector consisting of true component[1] and total flow rates in all streams of a process, the material balances and other process constraints can be written, at steady state, as

$$\boldsymbol{Bx} = \boldsymbol{0} \tag{2-1}$$

If chemical reactions occur, a master stoichiometric matrix $\boldsymbol{S}$ accounting for all reactions in the plant is so defined that the material balance equations are augmented to

[1] A component flow rate cannot be measured directly. It is rather obtained from the measurements of the corresponding total flow rate and concentration. For simplicity, we treat a component flow rate as if it were directly measured, if the corresponding total flow rate and concentration are both measured.

$$Bx + S^T \xi = 0 \tag{2-2}$$

where ξ is the vector of extents of reactions for all reaction units.

2.1.1 Categories of Variables

The variables in data reconciliation fall into four categories, as given in Table 2.1.

Table 2.1 Categories of Variables

Category	Notation	Explanation
0	c	Vector c consists of all exactly known component flow rates. The corresponding concentrations and total flow rates are exactly known and not adjustable.
1	$\tilde{x}$	Vector $\tilde{x}$ consists of all measured component flow rates. The corresponding concentrations and total flow rates are measured and adjustable.
2	$\tilde{d}$	Vector $\tilde{d}$ consists of measured concentrations and is adjustable. The corresponding unmeasured total flow rates are expressed by a diagonal matrix, N, which will be computed.
3	v	Vector v consists of all unmeasured component flow rates and the extents of reactions, ξ. The corresponding total flow rates are either measured or unknown.

The columns of matrix B are permuted and partitioned so that

$$B = [B_0 | B_1 | B_2 | B_3] \tag{2-3}$$

where columns of B_i correspond to the variables in category i (i = 0, 1, 2, 3). Matrix S is then combined with B_3 to form a new matrix

$$P = [B_3 | S^T] \tag{2-4}$$

which corresponds to the category 3 variables and the extents of reactions.

Energy balances can be added to the problem definition by considering enthalpy flow rate as the (C+1)st component.

2.1.2 Model of Steady State Data Reconciliation

Steady state data reconciliation was defined by Crowe (1986) as a constrained nonlinear programming problem

$$\mathcal{P}: \qquad \min_{a,\delta} F(a,\delta) = a^T \Sigma_1^{-1} a + \delta^T \Sigma_d^{-1} \delta \tag{2-5}$$

$$s.t.\ B_0 c + B_1(\tilde{x} + a) + B_2 N(\tilde{d} + \delta) + Pv = 0 \tag{2-6}$$

where a and δ are the vectors of adjustments to $\tilde{x}$ and $\tilde{d}$ so that Eq. (2-6) holds. Σ_1 and Σ_d are the variance-covariance matrices of $\tilde{x}$ and $\tilde{d}$, respectively. The process constraints expressed in Eq. (2-6) are known as the original process constraints.

In setting up the model of data reconciliation, $\mathcal{P}$, we assumed that $\tilde{x}$ and $\tilde{d}$ are mutually independent, and follow a certain probability distribution, $\tilde{x} \sim (x, \Sigma_1)$, $\tilde{d} \sim (d, \Sigma_d)$. This assumption usually holds, since correlations among category 1 variables $\tilde{x}$ and category 2 variables $\tilde{d}$ do not usually occur, unless correlations among streams cannot be ignored. A chemical reactor is such an example where variables in different streams are related by stoichiometry. A modification of $\mathcal{P}$ that takes account of such correlations is given later in this chapter.

2.2. Steady State Linear Data Reconciliation

2.2.1 Model and Solution

In Problem $\mathcal{P}$, if all total flow rates are measured and no category 2 variable is present, the bilinear term $N\delta$ and matrix B_2 in Eq. (2-6) will vanish. The Problem $\mathcal{P}$ is reduced to

$$\mathcal{P}_L: \quad \min_a \; F(a) = a^T \Sigma_1^{-1} a \tag{2-7}$$

$$s.t. \;\; B_0 c + B_1(\tilde{x} + a) + Pv = 0 \tag{2-8}$$

This is a quadratic programming problem having a unique solution for a since Σ_1 is positive definite. It is known as steady state linear data reconciliation and was studied by Crowe *et al.* (1983) and Crowe (1988, 1989a).

In order to remove the unknown category 3 variables, we define Y as a matrix of maximum rank whose columns span the null space of P^T, that is

$$Y^T P = 0 \tag{2-9}$$

The Problem $\mathcal{P}_L$ is then reduced to

$$\mathcal{P}_{L1}: \quad \min_a \; F(a) = a^T \Sigma_1^{-1} a \tag{2-10}$$

$$s.t. \;\; Y^T\left[B_0 c + B_1(\tilde{x} + a)\right] = 0 \tag{2-11}$$

The process constraints in Eq. (2-11) are known as the reduced process constraints. The procedure to remove the category 3 variables from the original constraints by Y is known as matrix projection. The matrix Y is not unique. A method to construct Y from P can be found in Crowe *et al.* (1983).

$\mathcal{P}_{L1}$ can be solved by Lagrange multipliers. Let us construct a Lagrange function

$$L(a,\lambda)=\frac{1}{2}\left(a^T\Sigma_1^{-1}a\right)-\lambda^T Y^T\left[B_0c+B_1(\tilde{x}+a)\right] \tag{2-12}$$

where λ is the vector of Lagrange multipliers for the reduced constraints. Naturally, the vector of Lagrange multipliers for the original constraints is

$$\mu=Y\lambda \tag{2-13}$$

The residuals of the reduced constraints is defined from Eq. (2-11) as

$$e=Y^T\left(B_0c+B_1\tilde{x}\right) \tag{2-14}$$

with the expectation of zeros and the variance-covariance matrix

$$H_e=cov(e)=Y^TH_0Y \tag{2-15}$$

where $H_0=B_1\Sigma_1B_1^T$, which is the variance-covariance matrix of e when there is no reaction nor category 3 variable in the process.

The optimal measurement adjustments a and the Lagrange multipliers λ can be obtained by finding a stationary point of the Lagrange function with respect to them. This gives

$$a=\Sigma_1B_1^TY\lambda \tag{2-16}$$

$$\lambda=-H_e^{-1}e \tag{2-17}$$

The expectation of a is a zero vector. The variance-covariance matrix of a is

$$Q_1=cov(a)=\Sigma_1B_1^TYH_e^{-1}Y^TB_1\Sigma_1 \tag{2-18}$$

The vector of the original residuals can be obtained from Eqs. (2-8) and (2-16) as

$$r = B_0 c + B_1 \tilde{x} + Pv = -B_1 a = H_0 Y H_e^{-1} e \tag{2-19}$$

with zero expectation and the variance-covariance matrix

$$H_r = cov(r) = H_0 Y H_e^{-1} Y^T H_0 \tag{2-20}$$

It can be shown that the optimal value of the objective function is

$$F = e^T H_e^{-1} e \tag{2-21}$$

The vector v can be obtained by solving Eq. (2-8).

The complete relationships among the variables in linear data reconciliation and the corresponding variance-covariance matrices are given in Appendix A. It is interesting to note that uniform expressions for both linear and bilinear data reconciliation exist for most of the variables in a compact form, which are highlighted in Appendix A and B.

The process constraints given in $\mathcal{P}_L$ do not impose the restriction that $\tilde{x} + a$ cannot be negative. Therefore, it is possible to get an infeasible adjustment to a measured component flow rate. Although the restriction can be easily implemented in a general nonlinear programming framework, we should point out that the infeasibility is usually caused by gross errors. When such errors are identified and removed, feasible and optimal adjustments should be obtained in most cases, without imposing any non-negativity constraint. Since the presence of an infeasible solution is itself an indication that a thorough inspection of the instrumentation and the process should be made, we do not recommend including any non-negativity constraint in the reconciliation model.

In general, bounds of variables can be treated within the optimization framework without difficulty[2]. However, whether or not to implement them is arguable. As some of

[2] Non-negativity constraints are of course bounds of variables.

the reconciled variables take values at their bounds, they distort hypothesis testing because the assumed statistical distribution for the variables breaks down at the bounds, unless a truncated distribution is justified and used[3]. We would recommend not including bounds in the reconciliation model unless they are used for diagnosis. Some references on this topic are Narasimhan and Harikumar (1993), and Crowe (1995).

2.2.2 Rank of the Variance-Covariance Matrices

The residuals e, r and the adjustments a are of great importance in hypothesis testing for gross errors. We now prove a fundamental relationship among them, the rank of H_r and the rank of Q_1 are both equal to the rank of H_e, i.e.,

$$rank(Q_1) = rank(H_r) = rank(H_e) \tag{2-22}$$

where H_e is nonsingular.

It is known that the rank of a product of matrices cannot be larger than the rank of any matrix in that product. Therefore $rank(Q_1) \le rank(H_e)$, in light of Eq. (2-18). On the other hand, we know from Appendix A that H_e can be expressed in terms of Q_1, $H_e = Y^T B_1 Q_1 B_1^T Y$, which suggests that $rank(H_e) \le rank(Q_1)$. This proves that $rank(Q_1) = rank(H_e)$.

The same argument applies to H_r. Because the expression of H_r involves H_e, according to Eq. (2-20), we have $rank(H_r) \le rank(H_e)$. On the other hand, it is shown in Appendix A that $H_e = Y^T H_r Y$, so that $rank(H_e) \le rank(H_r)$. This gives us that $rank(H_r) = rank(H_e)$.

[3] Under the truncated distribution, the probability is zero for a variable having a value beyond its bound.

It was known that Q_1 is always singular and H_r is singular as far as $H_r \neq H_e$. These are implied in Eq. (2-22) by considering the sizes of Q_1 and H_r.

2.3. Steady State Bilinear Data Reconciliation

A steady state bilinear data reconciliation problem was defined in Problem $\mathcal{P}$. It is called a bilinear data reconciliation because of the product term, $N\delta$, in the constraints. Both N and δ have to be computed in data reconciliation.

2.3.1 Model and Solution

Problem $\mathcal{P}$ involves an unknown matrix, N, which leads to a quadratic programming problem for any fixed N. An optimal solution can only be achieved through numerical iterations on N.

Crowe (1986) developed an algorithm to provide a sub-optimal solution to this problem. He later introduced a modified algorithm (Crowe, 1989b, 1994) that gives the optimal solution.

We mentioned earlier that the set-up of Problem $\mathcal{P}$ implies that there is no correlation between category 1 and 2 variables, which means that there is no correlation among different streams. A modification of Problem $\mathcal{P}$ that deals with a process of any correlation structure can be defined as

$$\mathcal{P}': \qquad \min_{a,\delta} F(a,\delta) = \begin{bmatrix} a^T & \delta^T \end{bmatrix} \Sigma^{-1} \begin{bmatrix} a \\ \delta \end{bmatrix} \tag{2-23}$$

$$s.t.\ B_0 c + B_1(\tilde{x} + a) + B_2 N(\tilde{d} + \delta) + Pv = 0 \tag{2-24}$$

where

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{1d} \\ \Sigma_{1d}^T & \Sigma_d \end{bmatrix} \tag{2-25}$$

is the variance-covariance matrix of $\begin{bmatrix} \tilde{x} \\ \tilde{d} \end{bmatrix}$. Σ_1 is the variance-covariance matrix of measured component flow rates (category 1 variables), Σ_d is the variance-covariance matrix of measured concentrations (category 2 variables), and Σ_{1d} is a matrix consisting of the covariances among the measured component flow rates and concentrations (among category 1 and category 2 variables). If no correlation exists between category 1 and category 2 variables, Σ_{1d} would be a zero matrix. Problem $\mathcal{P}'$ is then reduced to problem $\mathcal{P}$.

We now give the solution of Problem $\mathcal{P}'$. At the kth iteration, let the estimate of the measured concentrations be $\hat{d}^k = \tilde{d} + \delta^k$, where δ^k is the vector of adjustments to the measured concentrations at the kth iteration. The original constraints, Eq. (2-24), can be written as

$$B_0 c + B_1\left(\tilde{x} + a^k\right) + B_2 N^{k-1}\left(\hat{d}^{k-1} - \delta^{k-1} + \delta^k\right) + Pv = 0, \qquad k = 1, 2, \cdots \tag{2-26}$$

The optimally adjusted measurements and concentrations will be obtained at convergence

$$\hat{x} = \tilde{x} + \lim_{k\to\infty} a^k \tag{2-27}$$

$$\hat{d} = \tilde{d} + \lim_{k\to\infty} \delta^k \tag{2-28}$$

Analogous to the matrix Y defined in Eq. (2-9), another matrix Z of maximum rank whose columns span the null space of D^T is defined by

$$Z^T D = 0 \tag{2-29}$$

where D is defined with the jth column containing the columns of B_2 for stream j, multiplied by the concentrations, d_j, in that stream, namely

$$D[d] = Y^T \begin{bmatrix} B_{21}d_1 & B_{22}d_2 & \cdots & B_{2j}d_j & \cdots \end{bmatrix} \tag{2-30}$$

When the domain of Z and D, $[d]$, is emphasised, Eq. (2-29) is written as

$$\left(Z[d]\right)^T D[d] = 0 \tag{2-31}$$

Taking into account that $Y^T B_2 N\hat{d} = D\left[\hat{d}\right]n$ and $\left(Z\left[\hat{d}\right]\right)^T D\left[\hat{d}\right] = 0$, we can reduce the original constraints, Eq. (2-26), to

$$Z\left[\hat{d}^{k-1}\right]^T Y^T \left[B_0 c + B_1\left(\tilde{x} + a^k\right) + B_2 N^{k-1}\left(\delta^k - \delta^{k-1}\right)\right] = 0 \tag{2-32}$$

For simplicity, we denote $Z^k = Z\left[\hat{d}^k\right]$ and $Z = \lim_{k\to\infty} Z^k$. Problem $\mathcal{P}'$ is simplified to

$$\mathcal{P}_B: \quad \lim_{k\to\infty} \mathcal{P}'_B$$

where

$$\mathcal{P}'_B: \quad \min_{a^k,\delta^k} F\left(a^k,\delta^k\right) = \begin{bmatrix} a^{kT} & \delta^{kT} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} a^k \\ \delta^k \end{bmatrix} \tag{2-33}$$

$$s.t. \left(Z^{k-1}\right)^T Y^T \left[B_0 c + B_1\left(\tilde{x} + a^k\right) + B_2 N^{k-1}\left(\delta^k - \delta^{k-1}\right)\right] = 0 \tag{2-34}$$

where Z^{k-1}, N^{k-1}, and δ^{k-1} are available at the kth iteration.

We construct a Lagrange function $L = L\left(a^k, \delta^k, \lambda^k\right)$, and omit the superscript k for simplicity

$$L(a,\delta,\lambda) = \frac{1}{2}F - \lambda^T\left(Z^{k-1}\right)^T Y^T\left[B_0 c + B_1(\tilde{x}+a) + B_2 N^{k-1}\left(\delta - \delta^{k-1}\right)\right] \tag{2-35}$$

where λ is the vector of Lagrange multipliers as in the linear case for the reduced constraints. The vector of Lagrange multipliers for the original constraints is

$$\mu = YZ^{k-1}\lambda \tag{2-36}$$

Let

$$\Sigma^{-1} = \begin{bmatrix} \Sigma^1 & \Sigma^{1d} \\ \left(\Sigma^{1d}\right)^T & \Sigma^d \end{bmatrix} \tag{2-37}$$

Taking partial derivatives with respect to a and δ

$$\frac{\partial}{\partial a} L(a,\delta,\lambda) = \Sigma^1 a + \Sigma^{1d}\delta - B_1^T YZ^{k-1}\lambda = 0 \tag{2-38}$$

$$\frac{\partial}{\partial \delta} L(a,\delta,\lambda) = \left(\Sigma^{1d}\right)^T a + \Sigma^d \delta - N^{k-1}B_2^T YZ^{k-1}\lambda = 0 \tag{2-39}$$

and combining them to yield

$$\Sigma^{-1}\begin{bmatrix} a \\ \delta \end{bmatrix} = \begin{bmatrix} B_1^T \\ N^{k-1}B_2^T \end{bmatrix} YZ^{k-1}\lambda \tag{2-40}$$

which gives

$$\begin{bmatrix} a \\ \delta \end{bmatrix} = \Sigma\begin{bmatrix} B_1^T \\ N^{k-1}B_2^T \end{bmatrix} YZ^{k-1}\lambda \tag{2-41}$$

When Σ_{1d} is zero, it reduces to

$$a = \Sigma_1 B_1^T Y Z^{k-1} \lambda \tag{2-42}$$

$$\delta = \Sigma_d N^{k-1} B_2^T Y Z^{k-1} \lambda \tag{2-43}$$

which are the solutions given by Crowe (1994) to Problem $\mathcal{P}$.

The vector of the original residuals is defined, from Eq. (2-24), by

$$r = B_0 c + B_1 \tilde{x} + B_2 N \tilde{d} + Pv = -\left(B_1 a + B_2 N \delta\right) \tag{2-44}$$

and the vector of the reduced residuals are then defined by

$$e = Z^T Y^T \left(B_0 c + B_1 \tilde{x} + B_2 N \tilde{d}\right) = Z^T Y^T r \tag{2-45}$$

It can be shown that the expectations of e, r, a and δ are all zeros.

Let us denote $H_0 = \begin{bmatrix} B_1 & B_2 N \end{bmatrix} \Sigma \begin{bmatrix} B_1^T \\ N B_2^T \end{bmatrix}$, the variance-covariance matrices of r and e can be obtained from Eqs. (2-44) and (2-45), respectively

$$H_r = H_0 Y Z H_e^{-1} Z^T Y^T H_0 \tag{2-46}$$

$$H_e = Z^T Y^T H_0 Y Z \tag{2-47}$$

which are similar to Eqs. (2-20) and (2-15) in linear data reconciliation.

The variance-covariance matrix of $\begin{bmatrix} a \\ \delta \end{bmatrix}$ is expressed by

$$Q = \Sigma \begin{bmatrix} B_1^T \\ N B_2^T \end{bmatrix} Y Z H_e^{-1} Z^T Y^T \begin{bmatrix} B_1 & B_2 N \end{bmatrix} \Sigma \tag{2-48}$$

If we partition it as

$$Q = \begin{bmatrix} Q_1^* & Q_{1d} \\ Q_{1d}^T & Q_d^* \end{bmatrix} \tag{2-49}$$

we will have, when $\Sigma_{1d} = 0$, that

$$Q_1^* = Q_1 \tag{2-50}$$

$$Q_d^* = Q_d \tag{2-51}$$

where

$$Q_1 = \Sigma_1 B_1^T YZH_e^{-1} Z^T Y^T B_1 \Sigma_1 \tag{2-52}$$

$$Q_d = \Sigma_d N B_2^T YZH_e^{-1} Z^T Y^T B_2 N \Sigma_d \tag{2-53}$$

which are the variance-covariance matrices of a and δ of Problem $\mathcal{P}$, respectively, given by Crowe (1994).

It can be shown that the Lagrange multipliers are still expressed by

$$\lambda = -H_e^{-1} e \tag{2-54}$$

The vector of the unmeasured total flow rates, n, is updated by solving the following equations in the least square sense

$$D\left[\hat{d}\right] n = -Y^T \left(B_0 c + B_1 \left(\tilde{x} + a\right)\right) \tag{2-55}$$

The optimal value of the objective function can still be expressed by Eq. (2-21).

We note that Eq. (2-26) can also be written as

$$B_0 c + B_1(\tilde{x} + a) + B_2 N^{k-1}\hat{d} + Pv = 0, \qquad k = 1, 2, \cdots \tag{2-56}$$

therefore the partial derivatives of L with respect to n are

$$\left(D\left[\hat{d}\right]\right)^T Z^{k-1}\lambda = 0 \tag{2-57}$$

This condition holds approximately at each and every iteration, according to Eq. (2-31), and exactly when the algorithm converges as $Z^{k-1} \to Z^k$.

The comprehensive relationships among the variables in bilinear data reconciliation and the corresponding variance-covariance matrices are shown in Appendix B. The uniform expressions for both linear and bilinear data reconciliation are highlighted there.

The initial guess of the adjustments, δ^0, need not be feasible to satisfy process constraints. It can be conveniently chosen as $\delta^0 = 0$. Such a choice would virtually have no effect on the algorithm, because the algorithm will make the succeeding δ^k, $k = 1, 2, \cdots$, feasible anyway.

From Eq. (2-56), we may have

$$Y^T\left(B_0 c + B_1\tilde{x} + B_2 N\tilde{d}\right) = Y^T\left(B_0 c + B_1\tilde{x}\right) + D\left[\tilde{d}\right]n^0 = 0 \tag{2-58}$$

therefore, the initial guess of the unmeasured total flow rates can be made approximately feasible by solving the following equations in the least square sense

$$D\left[\tilde{d}\right]n^0 = -Y^T\left(B_0 c + B_1\tilde{x}\right) \tag{2-59}$$

The algorithm requires the computation of Z^k and D^k at each iteration. A method to obtain Z^k was given in Crowe (1986) based on the partitioning of D at the kth iteration

$$D = \begin{bmatrix} D_1 & D_3 \\ D_2 & D_4 \end{bmatrix} \tag{2-60}$$

where D_1 is a maximal square nonsingular submatrix in D.

Such partitioning is the most time consuming procedure in the algorithm. When programmed, it is much faster to verify if D_1 has a full rank than actually to identify a basis of D to form D_1. From Eq. (2-30) we know that

$$D\left[\hat{d}\right] = Y^T\left[B_{21}\hat{d}_1 \quad B_{22}\hat{d}_2 \quad \cdots \quad B_{2j}\hat{d}_j \quad \cdots\right] \tag{2-61}$$

where B_{2j} is constant and $\hat{d}_j$ is stochastic, $j = 1, 2, \cdots$. The partition structure of D is mainly determined by the constant matrix B_2. The vector $\hat{d}$ is unlikely to alter the partition because $\hat{d}$ is stochastic. In all the cases that we examined, the operation of extracting D_1 from D only needs to be done once. This significantly saves the computational time.

The vector v can be obtained by solving Eq. (2-24).

Data reconciliation is not complete and possibly not valid without computing v, provided that some of its elements are observable. We will demonstrate in a numerical example in Chapter 5 that a result of reconciliation given in literature was improper because an observable category 3 variable was negative.

2.3.2 Rank of the Variance-Covariance Matrices

As in the linear case, we now prove that the rank of H_r and the rank of Q are equal to the rank of H_e, i.e.

$$rank(Q) = rank(H_r) = rank(H_e) \tag{2-62}$$

while the rank of Q_1 and the rank of Q_d are less than or equal to the rank of H_e, i.e.

$$rank(Q_1) \le rank(H_e) \tag{2-63}$$
$$rank(Q_d) \le rank(H_e) \tag{2-64}$$

From Eq. (2-46) we know that H_r is calculated in terms of H_e, therefore $rank(H_r) \le rank(H_e)$, for the rank of a product of matrices cannot be larger than the rank of any matrix in that product. On the other hand, it can be seen from Appendix B that

$$H_e = Z^T Y^T H_r Y Z \tag{2-65}$$

thus $rank(H_e) \le rank(H_r)$. This proves that $rank(H_r) = rank(H_e)$.

The variance-covariance matrix of a, given in Eq. (2-52), is

$$Q_1 = \Sigma_1 B_1^T Y Z H_e^{-1} Z^T Y^T B_1 \Sigma_1 \tag{2-52}$$

which implies $rank(Q_1) \le rank(H_e)$.

The variance-covariance matrix of δ, given in Eq. (2-53), is

$$Q_d = \Sigma_d N B_2^T Y Z H_e^{-1} Z^T Y^T B_2 N \Sigma_d \tag{2-53}$$

which gives $rank(Q_d) \le rank(H_e)$.

The variance-covariance matrix of $\begin{bmatrix} a \\ \delta \end{bmatrix}$ is given by Eq. (2-48), It is obvious that $rank(Q) \le rank(H_e)$. However, from Eq. (2-48) we may have

$$H_r = (B_1 \quad B_2 N) Q \begin{pmatrix} B_1^T \\ N B_2^T \end{pmatrix} \tag{2-66}$$

therefore, $rank(H_e) = rank(H_r) \le rank(Q)$. This indicates that $rank(Q) = rank(H_e)$.

As in the linear problem, Q_1, Q_d, and Q are always singular. H_r is singular as far as $H_r \neq H_e$. These results are implied in Eqs. (2-62), (2-63), and (2-64).

2.4. Other Methods of Data Reconciliation

We pointed out that data reconciliation is indeed a constrained nonlinear programming problem. Besides the use of Lagrange multipliers to solve the problem, methods of constrained nonlinear programming can also be used. Veverka (1992) solved a nonlinear data reconciliation problem by successive linearization of nonlinear constraints. Islam *et al.* (1994) performed reconciliation to an industrial pyrolysis reactor based on simplified nonlinear mass and energy balances using Successive Linearization and Successive Quadratic Programming (SQP). Tjoa and Biegler (1991) presented an alternative data reconciliation and gross error detection approach for nonlinear systems using bivariate distribution function, where a hybrid SQP method was used. Schraa (1995) solved the Problem $\mathcal{P}$ by transferring the constrained problem into an unconstrained optimization problem.

If a general nonlinear programming code is used, extra computations have to be done in order to obtain the quantities used in hypothesis testing. If Lagrange multipliers

are used, as is the case in this chapter, those quantities are obtained simultaneously while data reconciliation is performed.

CHAPTER 3

DETECTING GROSS ERRORS

Measured process data are inherently inaccurate and violate process constraints because of their underlying stochastic behavior and of possible gross errors caused by process disturbances, process leaks, malfunctioning or miscalibrated instrumentation, or even departure from steady state.

We presented in Chapter 2 the models and the solutions for steady state data reconciliation. However, the optimal adjustments to measurements can be obtained if and only if there is no gross error. Therefore, all gross errors must be detected. They should then be corrected or the responsible measurements have to be deleted.

In this chapter, a review of the currently used statistical tests and a modification of the maximum power measurement test for gross errors are given. We will postpone our discussion about the advantages and the disadvantages of the tests until the principal component tests have been developed in the next chapter.

Data reconciliation is not complete and possibly not valid without statistical tests. We will point out in numerical examples in Chapter 5 that some reconciliations given in literature were improper and infeasible because they could not pass statistical tests.

It is also possible that no reconciliation would pass all of the statistical tests if the quality of measurements is so poor that there are too many gross errors. Under this circumstance, one should fix the instrumentation and possibly process leaks before reconciliation.

Several statistical tests have been developed for detecting gross errors. The chi-square collective test, first used in reconciliation by Reilly and Carpani (1963), compares the optimal value of the objective function in the model of data reconciliation to an appropriate tabulated chi-square value. They also proposed the univariate constraint test, which examines each residual of the process constraints. The univariate measurement test, which examines each measurement adjustment, was proposed by Mah and Tamhane (1982), and Crowe *et al.* (1983). Almasy and Sztano (1975) proposed a measurement test that possesses maximum power (MP) when there is only one gross error in the measurements. It is called the MP measurement test. The MP constraint tests were proposed by Crowe (1989b, 1992). General reviews of the statistical tests can be found in Crowe (1994), and Aldrich and Vandeventer (1994).

Though those statistical tests have been widely used for detecting gross errors, their performance has not always been satisfactory. Measurements, their variances and covariances, and the balance matrix together determine how the measurement adjustments should be made. Frequently, a subtle gross error that is too small to be detected by those tests could have a significant impact on quality and process control. When the tests fail to detect the error, the interpretation of plant performance may be distorted. In some cases, the tests will detect the presence of gross errors but will not correctly identify the variables in error. This is called confounding. Under such a circumstance, one may risk deleting good measurements and keeping corrupted ones. We will present a method to overcome these problems in the next chapter.

3.1. *A Priori* Knowledge about a Process

We already pointed out that measurements are contaminated by random errors and frequently by gross errors, the conservation laws and other process constraints are not satisfied. Generally this results in non-zero e, r, and a.

However, we know *a priori* that the expectation of e is always zero for measurements with only random errors. The balance residual vector e reflects the violation of the conservation laws and any other process constraints for which the measurements or any other process faults are responsible. On the other hand, H_e contains the information of the process structure, expressed in B, and of the measurement variance-covariance structure, expressed in Σ. H_e is the quantity that captures the process variability and the inherent process connectivity. These two quantities can help us in detecting gross errors that cannot be viewed as random events and violate the process covariance structure.

Since the projection matrix Y is not unique, there may be different sets of e and H_e. It is possible that a particular choice of Y leads to some small elements of e. Therefore e might not reveal all the gross errors in the process and measurements. Furthermore, when the process constraints are reduced by matrix projection, they no longer correspond directly to the physical process units in the process.

Fortunately, we also know *a priori* that the expectations of r and a are always zero for the measurements with only random errors. They do correspond to the original process units and measurements. We may examine them as well as e to find out if there is any gross error.

Matrix Σ is crucial in detecting gross errors and in data reconciliation, because the computation of all the other variables relies on the knowledge of Σ. Before samples are taken, all measuring instruments should be carefully checked and calibrated. The process should be inspected and any leaks should be fixed. The process then needs to be maintained at a steady state for a period of time during which repeated measurements are recorded. Care must be given to eliminate as many gross errors as possible before estimating Σ.

3.2. Univariate Tests

The univariate test of each reduced residual is given by

$$z_{e,i} = \frac{e_i}{\sqrt{(H_e)_{ii}}}, \qquad i = 1, 2, \cdots, m \tag{3-1}$$

where m is the rank of H_e.

Let us denote $diag(X)$ being a diagonal matrix whose diagonal elements are the same as those of X. We can write the univariate test statistics collectively

$$z_e = \left[diag(H_e)\right]^{-\frac{1}{2}} e \tag{3-2}$$

The univariate test of each original residual is given by

$$z_{r,i} = \frac{r_i}{\sqrt{(H_r)_{ii}}}, \qquad i = 1, 2, \cdots \tag{3-3}$$

or written collectively

$$z_r = \left[diag(H_r)\right]^{-\frac{1}{2}} r \tag{3-4}$$

The univariate test of each measurement adjustment is given by

$$z_{q,i} = \frac{q_i}{\sqrt{(Q)_{ii}}}, \qquad i = 1, 2, \cdots \tag{3-5}$$

and can be written collectively

$$z_q = \left[diag(Q)\right]^{-\frac{1}{2}} q \tag{3-6}$$

where $q = \begin{bmatrix} a \\ \delta \end{bmatrix}$. In a linear problem, δ vanishes. We have $q = a$ and $Q = Q_1$.

3.3. Maximum Power Tests

The MP test of each reduced residual is given by

$$z_{e,i}^* = \frac{\left(H_e^{-1} e\right)_i}{\sqrt{\left(H_e^{-1}\right)_{ii}}}, \qquad i = 1, 2, \cdots, m \tag{3-7}$$

or written collectively

$$z_e^* = \left[diag\left(H_e^{-1}\right)\right]^{-\frac{1}{2}} H_e^{-1} e \tag{3-8}$$

The MP test of each original residual is given by

$$z_{r,i}^* = \frac{\left(YZH_e^{-1} e\right)_i}{\sqrt{\left(YZH_e^{-1} Z^T Y^T\right)_{ii}}}, \qquad i = 1, 2, \cdots \tag{3-9}$$

or written collectively

$$z_r^* = \left[diag\left(YZH_e^{-1} Z^T Y^T\right)\right]^{-\frac{1}{2}} YZH_e^{-1} e \tag{3-10}$$

As pointed out by Crowe (1988, 1989b, 1992), the MP tests of the residuals $\boldsymbol{e}$ and $\boldsymbol{r}$ are identical to the tests of the Lagrange multipliers, λ and μ, respectively, because $\lambda = -\boldsymbol{H}_e^{-1}\boldsymbol{e}$ and $\mu = \boldsymbol{YZ}\lambda = -\boldsymbol{YZH}_e^{-1}\boldsymbol{Z}^T\boldsymbol{Y}^T\boldsymbol{r}$.

A modification of the MP test of each measurement adjustment is given by

$$z_{q,i}^{*} = \frac{\left(\Sigma^{-1}q\right)_i}{\sqrt{\left(\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix} YZH_e^{-1}Z^TY^T\begin{bmatrix} B_1 & B_2N \end{bmatrix}\right)_{ii}}}, \qquad i = 1, 2, \cdots \tag{3-11}$$

and can be written collectively

$$z_q^{*} = \left[diag\left(\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix} YZH_e^{-1}Z^TY^T\begin{bmatrix} B_1 & B_2N \end{bmatrix}\right)\right]^{-\frac{1}{2}} \Sigma^{-1}q \tag{3-12}$$

When Σ_{1d} in Eq. (2-25) is zero, Eq. (3-11) reduces to the MP measurement tests for $\boldsymbol{a}$ and δ given by Crowe (1992). Otherwise, the MP measurement tests should be performed on $\boldsymbol{q}$ instead of being carried out separately on $\boldsymbol{a}$ and δ.

If all covariances involving category 1 variables are zero, the MP test of $\boldsymbol{a}$ is identical to the univariate test, i.e., $z_a^{*} = z_a$. In this case, Σ_1 is diagonal. Similarly, if all covariances involving category 2 variables are zero, the MP test of δ is identical to the univariate test, i.e., $z_\delta^{*} = z_\delta$. In this case, Σ_d is diagonal.

3.4. A General Statistical Test

As a matter of fact, the statistics in Eqs. (3-2) and (3-8) are special cases of a general test statistic that examines any arbitrary linear combination of the reduced residuals

$$z_{ej}(V) = \frac{(Ve)_i}{\sqrt{\left(VH_eV^T\right)_{ii}}}, \qquad i = 1, \cdots, m \tag{3-13}$$

or written collectively

$$z_e(V) = \left[diag\left(VH_eV^T\right)\right]^{-\frac{1}{2}} Ve \tag{3-14}$$

It can be seen that z_{ej}, z_{ej}^* and $z_{ej}(V)$ are all unit normal variates if the measurements are normally distributed. The choice of V is arbitrary, except that it must be nonsingular. The univariate statistic has $V = I$, and the MP statistic has $V = H_e^{-1}$.

Similar arguments hold for the statistics given in Eqs. (3-4), (3-6), (3-10) and (3-12). For simplicity, we will focus our discussion on the statistics for the reduced residuals in the remaining of this and the next chapter. However, one should note that the analysis can be adapted to the statistics for the original residuals and the measurement adjustments.

3.5. Type I Error

The computation of the probability of a type I error for the conventional univariate and the MP tests is complicated due to the correlation among the residuals of process constraints. Conservative estimates are often used. One such estimate for that of an overall type I error is given by Sidak (1967)

$$\alpha = 1-(1-\alpha^*)^m \tag{3-15}$$

where α^* is the probability of a type I error for one of the residuals of the constraints. It is assumed that all residuals are subject to the same level of the type I error.

The conservative estimate of a type I error for e_i can be similarly obtained if the overall type I error, α, is given

$$\alpha^* = 1-(1-\alpha)^{\frac{1}{m}} \approx \frac{\alpha}{m} \tag{3-16}$$

The $\approx$ expression, which leads to the Bonferroni bound (Seber, 1984), holds when $\frac{\alpha}{m} << 1$.

3.6. Chi-Square Test

The global chi-square test is defined with m degrees of freedom by

$$\chi_m^2 = e^T H_e^{-1} e \tag{3-17}$$

which examines all the reduced residuals together. It is the optimal value of any objective function in the models of data reconciliation given in Chapter 2.

3.7. Identifying Gross Errors

Correctly detecting gross errors in a large-scale process is not an easy task. Nevertheless, it is even more difficult to identify them. Conventionally, gross errors are deemed to be detected if any of the statistics is larger than its threshold. Either the measurement or the process constraint that corresponds to the outlier statistic is regarded as most likely to be in gross error. Using this approach with the univariate and MP tests

may cause certain problems, as detailed in Chapter 5. Furthermore, identifying gross errors based on the chi-square test is difficult because $\boldsymbol{H}_{\boldsymbol{e}}^{-1}$ is not diagonal in general. In the next chapter, we will use principal component transformation to remove the inverses from the chi-square statistic. The contributing variables (e_i, $i = 1,\cdots,m$) are thus uncoupled from one another, which makes it easier to identify gross errors through the chi-square test.

CHAPTER 4

PRINCIPAL COMPONENT TESTS

In Chapter 3 we studied a number of currently used statistical tests for detecting gross errors. Unfortunately, they fail to give an alarm in some cases when there is a gross error, particularly when the gross error is subtle. Sometimes they do give an alarm, though there is no gross error at all, particularly in a large-scale process. Furthermore, they may give too many outliers when there are only a few gross errors. This phenomenon is called confounding. This may result in wrongly deleting good measurements while keeping corrupted ones.

We will derive a set of principal component (PC) tests in this chapter aimed at coping with those problems. The principal component tests are based on principal component analysis (PCA), which are sharper and less confounding in detecting and identifying gross errors. We will also compare the principal component tests to the other statistical tests, and explain why the former consistently perform better (sharper and less confounding) than the latter. Illustrations with numerical examples will be given in the next chapter.

PCA is an effective tool in multivariate data analysis. It transforms a set of correlated variables into a new set of uncorrelated variables, known as principal components (PCs). Each principal component is a linear combination of the original

variables. The coefficients of each linear combination are obtained from an eigenvector of the variance-covariance matrix of the original variables.

The idea of PCA was first introduced by Pearson (1901), and was generalized by Hotelling (1933). Good reviews of the theory of PCA can be found in Jolliffe (1986), Wold *et al.* (1987), and Jackson (1991). Kresta *et al.* (1991) presented a method of using PCA to monitor continuous processes, which Nomikos and MacGregor (1994) later extended to batch processes. A diagnostic method for finding out the causes of outliers by interrogating a PCA model was discussed by MacGregor *et al.* (1994). Tong and Crowe (1993) addressed the problems in developing principal component tests in on-line data reconciliation. They later developed a set of principal component tests used in steady state data reconciliation for detecting and identifying gross errors (Tong and Crowe, 1995).

4.1. PC Tests for Residuals of Reduced Process Constraints

For the sake of simplicity, we will focus our development of the principal component tests on the reduced residuals, e. At the end of this chapter, we will extend the method to the original residuals and measurement adjustments, which are directly related to the original process constraints, process units and measurements.

As seen in Chapter 3, the information in e can be analyzed by the univariate, MP, and chi-square tests. However, in order for us to gain a deeper insight into the problem, the analysis can be done differently. We will define a principal component transformation of the reduced residuals, and the tests that examine the principal components both individually and collectively.

4.1.1 Principal Component Transformation

Let us consider a set of linear combinations of e

$$y_e = W_e^T\left(e - e^*\right) = W_e^T e \tag{4-1}$$

where e^* is the expectation of e that is a zero vector in data reconciliation.

If the linear combination coefficients given in W_e form eigenvectors of H_e, vector y_e will be principal components with the values being principal component scores. To ensure the uniqueness of W_e, we may restrict

$$W_e = U_e \Lambda_e^{-\frac{1}{2}} \tag{4-2}$$

Matrix Λ_e is diagonal, consisting of the eigenvalues of H_e, $\lambda_{e,i}$, $i = 1, \cdots, m$, on its diagonal, and satisfies

$$\Lambda_e = U_e^T H_e U_e \tag{4-3}$$

Matrix U_e consists of the orthonormalized eigenvectors of H_e, so that

$$U_e U_e^T = I \tag{4-4}$$

It can be shown that $y_e \sim (0, I)$ because $e \sim (0, H_e)$. Therefore, a set of correlated variables, e, is transformed into a new set of uncorrelated variables, y_e, with unit variances. There is no covariance among the principal components. The elements (PCs) of y_e are in descending order according to the magnitudes of the corresponding eigenvalues of H_e.

On the other hand, Eqs. (4-1) and (4-2) can be combined and rewritten as

$$e = e^* + U_e \Lambda_e^{\frac{1}{2}} y_e \tag{4-5}$$

This means that the residual vector e can be uniquely reconstructed from its principal components if all of the principal components are retained, i.e., $y_e \in R^m$. However, if fewer than m of them are retained, we will have

$$e = e^* + U_e \Lambda_e^{\frac{1}{2}} y_e + (e - \hat{e}) \tag{4-6}$$

with

$$\hat{e} = e^* + U_e \Lambda_e^{\frac{1}{2}} y_e \tag{4-7}$$

where $y_e \in R^{k_e}$, and $k_e < m$. Eq. (4-7) is referred to as the principal component model of e. Eq. (4-6) indicates that the residuals in the vector e can be decomposed into the contributions from their expectations, e^*, principal components, y_e, and the residuals of the principal component model, $e - \hat{e}$.

If the measured variables are normally distributed, $\begin{bmatrix} \tilde{x} \\ \tilde{d} \end{bmatrix} \sim N(\begin{bmatrix} x \\ d \end{bmatrix}, \Sigma)$, we have, from our earlier discussion, that

$$y_e \sim N(0, I) \tag{4-8}$$

Instead of looking at statistical tests of e, we may study alternatively how to perform hypothesis testing on y_e and $e - \hat{e}$. We will show that when each PC is tested individually, all of the PCs must be examined one by one. However, when they are tested collectively, there are two tests where one examines the retained PCs collectively and the other examines the unretained ones collectively[1].

[1] The retained PCs are those which can approximate the process under normal operating conditions, according to certain criterion (a stopping rule).

4.1.2 Principal Component Test

Based on Eqs. (4-1) and (4-8), the test statistic of a principal component is defined as

$$y_{e,i} = (W_e^T e)_i \sim N(0,1), \qquad i = 1,\cdots,k_e \tag{4-9}$$

which can be tested against a threshold tabulated value.

Eq. (4-9) shows that the *i*th principal component, $y_{e,i}$, is obtained from the inner product of e and the *i*th eigenvector of W_e.

Let us recall the general test statistic given in Eq. (3-13). If we substitute W_e^T for V and take into account that $W_e W_e^T = H_e^{-1}$, we can see that the PC test shown in Eq. (4-9) is a special case of the general test statistic.

4.1.2.1 Contribution Analysis

Once a gross error is detected, it has to be located. This is called gross error identification.

We can identify the constraints in gross error by inspecting the contribution from the *j*th residual, e_j, to a suspect principal component, say $y_{e,i}$, which is calculated by

$$g_j = \left(w_{e,i}\right)_j e_j, \qquad j = 1,\cdots,m \tag{4-10}$$

Define $g = \left(g_1,\cdots,g_m\right)^T$, and let g' be the same as g except that its elements are sorted in descending order based on their absolute values. We can study the contributions by checking the signs and the magnitudes of the elements in g'. In general, the

contributions will vary, and are dominated by the first few elements. They are major contributors to the suspect principal component, and are directly related to the constraints that should also be suspected. The number of the major contributors, k_1, can be set so that

$$\left| \frac{\left(\sum_{j=1}^{k_1} g'_j \right) - y_{e,i}}{y_{e,i}} \right| \leq \varepsilon_1 \tag{4-11}$$

where ε_1 is a prescribed tolerance.

It is noted that since the signs of these contributions can be either plus or minus as can the signs of the elements of $w_{e,i}$ and e, the cancellation effect among the elements of g' should be taken into account in identifying the suspect constraints. This is done in Eq. (4-11).

4.1.2.2 Type I Error

When $k_e \leq m$ principal components are retained, Eqs. (3-15) and (3-16) can be rewritten as

$$\alpha = 1 - (1 - \alpha^*)^{k_e} \tag{4-12}$$

$$\alpha^* = 1 - (1 - \alpha)^{\frac{1}{k_e}} \approx \frac{\alpha}{k_e} \tag{4-13}$$

the $\approx$ expression holds when $\frac{\alpha}{k_e} << 1$.

Eq. (4-12) can be used to calculate the *exact* probability of the overall type I error from a prescribed α^* for the principal components, and Eq. (4-13) can be used to

calculate the *exact* probability of a type I error for a principal component from a prescribed overall type I error, α, because the principal components are not correlated.

4.1.3 Collective Tests

4.1.3.1 Truncated Chi-Square Test

Analysis of the causes that inflate the chi-square test, Eq. (3-17), is difficult because of the presence of H_e^{-1} that confounds the contributions from the elements of e to χ_m^2. However, the principal components can be related to the chi-square statistic through

$$\chi_m^2 = e^T H_e^{-1} e = y_e^T y_e \tag{4-14}$$

where $y_e \in R^m$. If $k_e < m$ principal components are retained, i.e., $y_e \in R^{k_e}$, we have

$$\chi_{k_e}^2 = y_e^T y_e \tag{4-15}$$

which is called the truncated chi-square test, and is a principal component approximation of the chi-square statistic.

By employing the principal components, as expressed in Eqs. (4-14) and (4-15), we can identify the constraints in gross error by checking the magnitudes of the retained principal components. Those principal components having large absolute scores are major contributors to the inflated statistic, and are flagged as suspects. It should be noted that the principal components that are major contributors to a chi-square statistic are not necessarily outliers themselves. It is the residuals of the constraints that significantly contribute to the principal components that need to be studied further. Again, this can be done by checking the elements of g'. Eq. (4-11) can still be used to determine the number of the major contributors from the constraints.

It should be noted that the expression of the chi-square statistic in terms of the principal components allows us not only to detect but also to identify the constraints in gross error from that statistic, which is impossible otherwise.

4.1.3.2 Q_e Test

Another important collective test statistic is defined by

$$Q_e = (e-\hat{e})^T (e-\hat{e}) \tag{4-16}$$

It is the squared prediction error, known also as the Q statistic or the Rao-statistic. It can be shown that

$$Q_e = \sum_{i=k_e+1}^{m} \lambda_{e,i} y_{e,i}^2 \tag{4-17}$$

hence Q_e is a weighted sum of squares of the last $m-k_e$ principal components. It is a quadratic form, and is a linear combination of chi-square variables of one degree of freedom.

If we define $q_1 = \sum_{i=k_e+1}^{m} \lambda_{e,i}$, $q_2 = \sum_{i=k_e+1}^{m} \lambda_{e,i}^2$, $q_3 = \sum_{i=k_e+1}^{m} \lambda_{e,i}^3$, and $h_0 = 1 - \frac{2q_1 q_3}{3q_2^2}$,

the upper threshold for Q_e can be obtained as

$$Q_{e,\alpha} = q_1 \left(\frac{c_\alpha \sqrt{2q_2 h_0^2}}{q_1} + \frac{q_2 h_0 (h_0 - 1)}{q_1^2} + 1 \right)^{\frac{1}{h_0}} \tag{4-18}$$

where c_α is a one-tail threshold value of unit normal variate subject to $c_\alpha h_0 < 0$ (Jackson, 1991).

4.1.3.3 Contribution Analysis

To analyze the causes that inflate the Q_e statistic, similar to the analysis of $\chi^2_{k_e}$, we look at the absolute values of the residuals of the predictions, i.e., the elements of $e-\hat{e}$, sorted in descending order. Those elements having large absolute values are major contributors to the inflated Q_e, and are flagged as suspects. Again, the residuals that are major contributors to Q_e are not necessarily outliers themselves, because the univariate statistic on e_i often cannot pick up an outlier when residuals are highly correlated, as illustrated in the next section.

Let $f = e-\hat{e}$, and f' be the same as f, except that its elements are sorted in descending order based on their absolute values. The number of the major contributors to Q_e, k_2, is given by

$$1-\frac{\sum_{i=1}^{k_2} f_i'^2}{Q_e} \le \varepsilon_2 \tag{4-19}$$

where ε_2 is a prescribed tolerance similar to ε_1.

Traditionally, we need to check all $k_e = m$ univariate statistics, each corresponding to a residual of the constraints. If we only looked at $k_e < m$ residuals, we would have no idea about the other $m-k_e$ residuals. However, since each principal component contains the information from all residuals, provided that $\boldsymbol{H}_e$ is not diagonal, we are able to choose $k_e < m$ principal components to represent the original problem, where k_e is problem dependent. It is reasonable to expect that these retained principal components would be able to pick up unusual events that inflate one or more residuals, as is usually the case. However, some events may not be detected by the retained principal

components. This may be because the gross errors are caused by process disturbances which have been ignored by the principal component model. Therefore, we need to look at the principal components that are not retained in the model, through the Q_e statistic.

4.1.3.4 Stopping Rules

Different stopping rules give quite different values of k_e, and a comprehensive review can be found in Jackson (1991). By appropriately choosing k_e, we can use the PC model to mainly take account of the process variability, while leave the inherent variability to the unretained principal components.

To choose such a k_e, we follow Horn (1965). Let H'_e be a diagonal matrix whose diagonal elements are the same as those of H_e, i.e., $H'_e = diag(H_e)$, $\lambda_e \in R^m$ be all of the eigenvalues of H_e, and $\lambda'_e \in R^m$ be all of the eigenvalues of H'_e, both in descending order. The recommended number of the principal components to be retained is given by

$$k_e = max(i) \left| \lambda_{e,i} \geq \lambda'_{e,i}, \qquad i = 1, \cdots, m \right. \tag{4-20}$$

In practice, $k_e << m$, especially when m is large.

The reason that we use Horn's rule to determine k_e is that we can obtain k_e directly from H_e without relying on any process data. We choose k_e based on Eq. (4-20), because when $\lambda'_{e,i} > \lambda_{e,i}$, the inherent variability becomes dominant, and that is where we should stop adding more principal components.

4.1.3.5 Relationship between $\chi^2_{k_e}$ and Q_e

$\chi^2_{k_e}$ and Q_e are complementary in that the former examines the retained and the latter examines the unretained principal components, collectively. $\chi^2_{k_e}$ accounts for the amount of variance explained by the principal component model, while Q_e accounts for the amount of the variance unexplained. Q_e will be inflated whenever an assignable cause is involved in this quantity.

4.1.3.6 Relationship between y_e and the Collective Tests $\chi^2_{k_e}$ and Q_e

Using PCA to detect gross errors in data reconciliation is not exactly the same as using it to approximate and interpret a process, though they use the same statistical tool. When PCA is used to interpret a process, as in Nomikos and MacGregor (1994), only the first few PCs are retained, since those PCs can explain major variances in data. However, when PCA is used in data reconciliation, all PCs have to be analyzed. We need to investigate each and every y_{ei} in testing individual PCs, and both $\chi^2_{k_e}$ and Q_e in testing the PCs collectively. This is because a gross error can violate any part of the process covariance structure.

4.2. Relationships Among the Statistical Tests

The matrix H_e is not diagonal in general even if Σ itself is, because H_e contains the information from the topology of the flowsheet, which correlates the measured variables. It is well known that as the correlation increases, the performance of the univariate test becomes less acceptable. We will see that the same argument almost equally

applies to the MP test. However, the performance of the PC test is consistent regardless of the correlation in the variables, as will be demonstrated later.

We first prove that when H_e is diagonal, the univariate, MP and PC constraint tests are all identical.

The general test statistic, defined in Eq. (3-14), reduces to

$$z_e(V) = \left[VH_eV\right]^{-\frac{1}{2}}Ve \tag{4-21}$$

when V, as well as H_e, is diagonal.

It is easy to show that the univariate and the MP statistics can be written as

$$z_e(V) = H_e^{-\frac{1}{2}}e \tag{4-22}$$

where $V = I$ or H_e^{-1}, respectively.

For the principal component test statistic defined by Eqs. (4-1) and (4-2), we have, when H_e is diagonal, $V = W_e^T = H_e^{-\frac{1}{2}}$, since $U_e = I$ and $\Lambda_e = H_e$. Therefore V is diagonal, and Eq. (4-1) also reduces to Eq. (4-22).

Example 5.1 illustrated a case where H_e is diagonal.

We see that the three types of statistics are all identical in the limit of diagonal H_e. Geometrically they would form three identical rectangular or hyper-rectangular regions to approximate the joint elliptical confidence region in the coordinates spanned by the vector e, and therefore would be indistinguishable, as shown in Figure 4.1. It means that when H_e is close to being diagonal, the performance of the three tests is very similar. However, H_e is diagonal only when there are no chemical or thermodynamic interactions of any kind in a system, and no unit in the system is connected to any other unit[2]. In this case, data reconciliation could be done on the individual components in each unit. Nevertheless, when H_e is diagonal, the univariate and the MP test are identical to the PC test, thus there is no need to perform PCA. The matrix H_e is not diagonal in general, and in fact, can differ greatly in different plants, thus affecting the quality and the reliability of the tests. To demonstrate, we look at a two dimensional problem where a graph can easily be drawn.

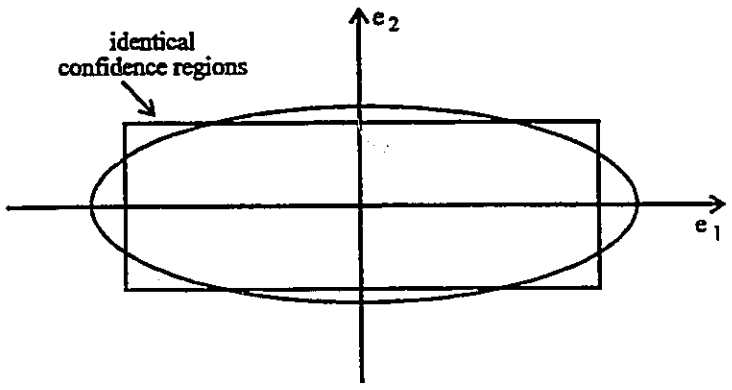


Figure 4.1 Identical Confidence Regions of the Univariate, MP, and PC Tests if H_e is Diagonal

As H_e shifts away from diagonal form, it can be shown that the confidence regions of the univariate and the PC tests remain rectangular, while the MP region changes gradually from a rectangle to a flattened parallelogram, as can be seen from Eqs. (3-2), (3-8), and (4-1), in comparison to Eq. (4-22). An example of the relative positions of the confidence regions of the tests is illustrated in Figure 4.2 along with the ellipse of the chi-

[2] Here a unit may be a physical unit or a "conceptual" unit that consists of lumped units.

square statistic, when H_e is not diagonal. The inadequacy of using the univariate region under such condition is well known and discussed in many statistical textbooks, while the inadequacy of using the MP region has not been addressed to our knowledge. In fact, the MP test may be the worst one in many cases among all the tests that we considered in this thesis, as also illustrated in Figure 4.2.

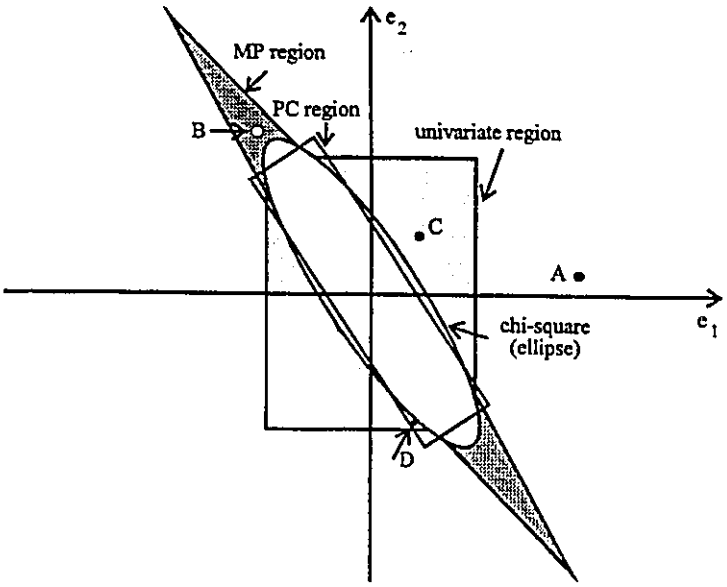


Figure 4.2 Relationship Among z_e, z_e^*, y_e, and χ_m^2 Tests When H_e is not Diagonal

We observe from Figure 4.2 that if there is only one gross error of large magnitude, such as that represented by the point A, all tests are able to pick up the unusual event. The choice of the tests does not matter. However, for example, when two relatively large gross errors, represented by the point B, happen to fall in the dark shaded area in the MP region, the conclusion drawn from the MP test would be in error. Similarly, if two gross errors, represented by the point C, happen to fall in the light shaded area, the conclusion drawn from the univariate test would be wrong. Due to the flattened shape of the MP region, it has a potential to wrongly accept large gross errors, such as the point B, which would be rejected by all the other tests. This is why we said that the MP test might be the worst one under certain circumstances. On the contrary, the relative position between the PC region and the ellipse is fixed under any chosen type I error regardless of the structure of H_e. This leads to the consistent performance of the PC test.

We know that the difference between the univariate and the χ^2_m tests is that the former does not take the correlation among the residuals into account and hence tends to be less reliable when correlation increases. This suggests that multivariate data should be tested against multivariate criteria. The MP test, on the other hand, does incorporate the correlation by including the inverse of the covariance matrix in its formula. This leads to its maximum power for correctly detecting a gross error over all the tests defined in Eq. (3-14), including the univariate and the PC tests, but only when there is a single error. When there are multiple gross errors, it will no longer possess the maximum power, as we have already seen, also due to its use of the inverse of the covariance matrix. The PC test takes the correlation into account by implementing the eigenvectors in its formula, thereby overcoming the drawbacks that the MP test exhibited.

As an illustration that the MP test possesses the maximum power when there is only one gross error, we look at the point D. The point is located outside of the MP region, but within the univariate and the PC regions. Figure 4.2 shows that the probability of having such a point is not very large. In most single gross error cases, even though the MP test has a higher power, it does not prevent the other tests from being able to detect that error.

In the numerical examples that we will present in the next chapter, we will show that the principal component tests not only provide better detection even for subtle gross errors, but also have a substantially greater power to correctly identify the variables in error over the other tests.

4.3. Test for the Original Constraints and the Adjustments

We have restricted ourselves to the reduced constraints for the sake of simplicity. However, if the vector of the reduced residuals, e, is replaced by the vector of the original residuals, r, and H_e is replaced by H_r, the test procedures that we presented previously are still valid. But H_r is of full rank (identical to H_e) only in the absence of unmeasured quantities, and is singular when unmeasured quantities or chemical reactions are present in a process. When it is singular, the maximum number of the principal components that can be retained, or equivalently, its rank, is less than the number of original constraints. This is because linear relationships exist among the reconciled variables or unmeasured quantities. In this case, the eigenvalues (sorted in descending order) indexed beyond the maximum number of retainable principal components are zero.

When e is replaced by a (linear problem) or $q = \begin{bmatrix} a \\ \delta \end{bmatrix}$ (bilinear problem), and H_e is replaced by Q_1 or Q, respectively, the test procedures are also valid. But as we know, Q_1 and Q are never of full rank. The maximum number of the principal components that can be retained, or equivalently, the ranks of Q_1 or Q, are less than the number of the measurements. The eigenvalues (sorted in descending order) indexed beyond the maximum number of retainable principal components are also zero.

It has been shown in Chapter 2 that H_e, H_r and Q_1 have the same rank in a linear problem, while H_e, H_r and Q have the same rank in a bilinear problem, therefore we restrict

$$k_r \leq \text{rank}(H_r) = \text{rank}(H_e) = m \tag{4-23}$$

$$k_a \le \text{rank}(Q_1) = \text{rank}(H_e) = m \tag{4-24}$$

$$k_q \le \text{rank}(Q) = \text{rank}(H_e) = m \tag{4-25}$$

where k_r, k_a and k_q are the numbers of principal components that may be retained in the corresponding tests. In general, k_e, k_r and k_a (or k_q) may be different in a problem.

In a bilinear problem, since the matrices H_e, H_r, and Q are not constant during reconciliation, the principal component tests should be employed only at the final stage of data reconciliation, where those covariance matrices have their final values.

4.4. Practical Issues in Performing the PC Tests

Principal component tests provide an insight into a correlated problem, which might not be obtained otherwise. The rule of thumb is that the collective test statistics should be examined first. When outliers of any of those statistics are observed, we may switch to the principal component graphs and the contribution graphs to identify the variables in gross error. The univariate and MP tests can still be used as additional evidence. Detailed examples will be given in the next chapter.

The principal component tests can be applied to any case with a known probability distribution of the measurements. When normal distribution could not be assumed, the average of measurements of a variable would tend toward a normal distribution as the size of the measurements increases because of the central limit theorem, provided that they have a finite expectation and variance, and are independent. However, since measurements are usually correlated in time, the central limit theorem may not hold. When the distribution is unknown, enough samples have to be taken to obtain a reference distribution.

Statistical tests work in the probability sense Regardless of the power of the principal component tests, there is no guarantee that they can always detect a gross error if there is one in the process. Therefore the principal component tests should not be used alone. This comment applies equally to all the other statistical testing methods. The principal component tests involve more computation in calculating eigenvalues and eigenvectors, and more analysis of contribution graphs. This is however not a practical problem with today's computers.

CHAPTER 5

NUMERICAL EXAMPLES

We have presented models of data reconciliation and a number of statistical tests in previous chapters. In this chapter, we will provide a number of examples to illustrate the methods. In particular, we will show that the PC tests are more sensitive to subtle gross errors, and have a substantially greater power to correctly identify the variables in error than the other tests[1,2].

5.1. A process with a leak

This example was used by Crowe *et al.* (1983) to demonstrate how to detect an unsuspected leak in a process. The process consists of three units in series, with an unknown leak in the second unit, as shown in Figure 5.1. Only total mass balances were considered. Because all flow rates were measured, the original constraints are identical to the reduced constraints, therefore, $H_e = H_r$, $e = r$. The matrices used in this example are

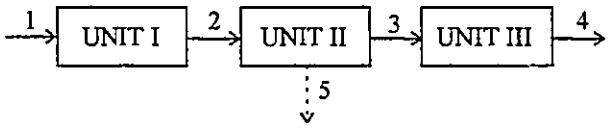


Figure 5.1 A Process with a Leak

[1] In each example, all the statistical tests mentioned in Chapter 3 and 4 were performed. However, some may not be shown to save space.

[2] All statistics were examined at 95% overall confidence level, unless otherwise stated.

$$A = B_1 = \begin{bmatrix} 1 & -1 & . & . \\ . & 1 & -1 & . \\ . & . & 1 & -1 \end{bmatrix}, \qquad Y^T = \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix}, \qquad \Sigma_1 = \begin{bmatrix} 1 & . & . & . \\ & 1 & . & . \\ & & 1 & . \\ sym. & & & 1 \end{bmatrix}$$

$$H_e = H_r = \begin{bmatrix} 2 & -1 & . \\ & 2 & -1 \\ sym. & & 2 \end{bmatrix}, \qquad Q_1 = \begin{bmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ & 0.75 & -0.25 & -0.25 \\ & & 0.75 & -0.25 \\ & sym. & & 0.75 \end{bmatrix}$$

The rank of the covariance matrices H_e and Q_1 is 3. Obviously Q_1 is singular because it does not have full rank. On the other hand, H_e is non-singular.

The true values of the total flow rates in streams 1 through 4 are $x = [100 \quad 100 \quad 95 \quad 95]^T$.

5.1.1 Case 1: Original Data

In this case, the measurements of total flow rates given in Crowe *et al.* (1983) were used. The leak is easily observable, because the residual of the material balance in unit II is 4.5, which is more than three times larger than its standard deviation, 1.41. We see from Tables 5.1, 5.2 and 5.3[3] that all tests were able to detect the leak. The threshold of a unit normal variate with 95% overall confidence and 3 degrees of freedom is 2.39. Outliers and major contributors to the outlier PCs are shaded in the tables.

With the use of the PCs, we are able to identify gross errors based on inflated chi-square and truncated chi-square statistics. The major contributors to χ^2_m are the first and

[3] The numbers given in the parentheses in the first columns of Tables 5.1, 5.2 and 5.3 are the equation numbers in the earlier chapters used to compute the quantities prior to them, respectively. Where the equation for computing a particular quantity was not provided in the earlier chapters, the equation number of a similar equation is given.

the third PCs, and the major contributor to $\chi^2_{k_e}$ is the first PC. It is seen from Table 5.2 that these PCs are primarily related to the flow rate balance around unit II. The contributors to $\chi^2_{k_a}$ are the first and the second PCs, which are primarily related to the measurements taken from unit II, as can be seen from Table 5.1.

Table 5.1 The Process with a Leak: Measurements and Measurement Tests

	1	2	3	4
$\tilde{x}$	98.5	101.0	96.5	95.5
$\hat{x}$; ($\tilde{x}$ + (2-16))	97.875	97.875	97.875	97.875
$z_a = z_a^*$; (3-6)	-0.722	**-3.608**	1.588	**2.742**
y_a; (4-1)	**2.733**	0.933	**3.057**	—
Contributions to $y_{a,1}$; (4-10)	0.089	**1.769**	**1.108**	-0.233
Contributions to $y_{a,2}$; (4-10)	-0.495	1.739	-0.343	0.033
Contributions to $y_{a,3}$; (4-10)	0.200	**1.082**	-0.268	**2.043**

Table 5.2 The Process with a Leak: Constraint Tests

	1	2	3
$r = e$; (2-14)	-2.5	**4.5**	1.0
$z_r = z_e$; (3-2)	-1.768	**3.182**	0.707
$z_r^* = z_e^*$; (3-8)	0.722	**3.750**	**2.742**
$y_r = y_e$; (4-1)	-2.128	-1.750	**3.178**
Contributions to $y_{e,1}$; (4-10)	-0.677	**-1.722**	0.271
Contributions to $y_{e,2}$; (4-10)	-1.250	0.000	-0.500
Contributions to $y_{e,3}$; (4-10)	-1.633	**-4.158**	0.653

Table 5.3 The Process with a Leak: Collective Tests

Statistic	Value	Threshold	Degrees of Freedom	Contributions
χ^2_m; (2-21)	**17.68**	7.82	$m = 3$	1st & 3rd PCs
$\chi^2_{k_e}$; (4-15)	**7.59**	5.99	$k_e = 2$	1st PC
Q_e; (4-16)	**5.91**	2.20	$k_e = 2$	see Table 5.4
$\chi^2_{k_a}$; (4-15)	**8.34**	5.99	$k_a = 2$	1st & 2nd PCs
Q_a; (4-16)	**9.35**	3.75	$k_a = 2$	see Table 5.4

Contribution analysis of Q_e also revealed that the leak in unit II was responsible for that statistic. However, the contribution analysis of Q_a pointed to the flow rate measurement in stream 4 because of confounding.

Table 5.4 The Process with a Leak: Contributions to the Q statistics

Contributions from e_i to Q_e			Contributions from a_j to Q_a			
$i=1$	$i=2$	$i=3$	$j=1$	$j=2$	$j=3$	$j=4$
1.479	2.957	1.479	0.954	1.120	0.354	6.918

5.1.2 Case 2: A Subtle Leak

In order to make the leak harder to detect, a different set of measurements was simulated. In this case, the residual of the balance in unit II becomes 2.1, which is only about 50% larger than its standard deviation. Though the leak was small, it violated the process variance-covariance structure. The leak passed all the non-PC tests.

Table 5.5 The Process with a Subtle Leak: Measurements and Reconciliation

	1	2	3	4
$\tilde{x}$	98.4	98.6	96.5	96.2
$\hat{x}$	97.425	97.425	97.425	97.425

Table 5.6 The Process with a Subtle Leak: Collective Tests

Statistic	Value	Threshold	Degrees of Freedom	Contribution
χ^2_m	4.688	7.82	$m=3$	—
$\chi^2_{k_e}$	0.666	5.99	$k_e=2$	—
Q_e	**2.356**	2.20	$k_e=2$	see Table 5.7
$\chi^2_{k_a}$	2.153	5.99	$k_a=2$	—
Q_a	2.535	3.75	$k_a=2$	—

As seen from Table 5.6, only the Q_e statistic was able to detect the gross error at 95% overall confidence level. This demonstrates the advantage of using the PC test to detect a subtle gross error. The second residual of the constraints contributes the most to Q_e, which is associated with the leak.

Table 5.7 The Process with a Subtle Leak: Contributions to Q_e statistic

Contributions from e_i to Q_e		
$i=1$	$i=2$	$i=3$
0.589	1.178	0.589

5.1.3 Case 3: The Leak is Accounted For

If we denote the leak as stream 5, and study the process again with the same measurements as in Case 2, there would be no outlier from any of the tests. The matrices given at the beginning of this example change to

$$A=\begin{bmatrix}1 & -1 & . & . & . \\ . & 1 & -1 & . & -1 \\ . & . & 1 & -1 & .\end{bmatrix} \qquad Y^T=\begin{bmatrix}1 & . & . \\ . & . & 1\end{bmatrix} \qquad H_e=\begin{bmatrix}2 & . \\ . & 2\end{bmatrix}$$

$$H_r=\begin{bmatrix}2 & -1 & . \\ & 1 & -1 \\ sym. & & 2\end{bmatrix} \qquad Q_1=\begin{bmatrix}0.5 & -0.5 & . & . \\ & 0.5 & . & . \\ & & 0.5 & -0.5 \\ sym. & & & 0.5\end{bmatrix} \qquad Y^TB_1=\begin{bmatrix}1 & -1 & . & . \\ . & . & 1 & -1\end{bmatrix}$$

The residual of the original balance around unit II is -0.05 in this case, which is negligible. The rank of the variance-covariance matrices is 2. Unit II is effectively removed from the process by the matrix projection. The first reduced constraint is the balance around unit I, and the second is the balance around unit III. The two parts of the process (unit I and III) are independent. The threshold for a uninormal variate with 95% overall confidence level and 2 degrees of freedom is 2.24. The PC test of e is exactly the same as

the univariate and MP tests, because H_e is diagonal. The detailed test results are omitted because none of the statistics was an outlier.

This problem gives an example that H_e is diagonal. Of course it is so because of our choice of Y^T. Since Y^T is not unique, we may alternatively choose

$$Y^T = \begin{bmatrix} 1 & . & . \\ 1 & . & 1 \end{bmatrix}$$

which gives

$$Y^T B_1 = \begin{bmatrix} 1 & -1 & . & . \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

and H_e is no more diagonal since

$$H_e = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$

The statistics for e will change values, but all tests would be passed anyway.

A closer look at the reduced balances reveals that the first reduced constraint is the balance around unit I, and the second is the balance around the lumped units I and III. The two parts of the process (unit I and III) are still independent. This indicates that the data reconciliation can be done independently on each unit, and there is no need to perform the PC test since the univariate and MP test would be identical to the PC test. However, if H_e is used as given, the PC test should be done since H_e is not diagonal.

5.2. Sunoco Hydrocracker Fractionation Plant

The Sunoco hydrocraker fractionation plant studied by Bailey (1991) is shown in Figure 5.2. The hydrocraker plant converts gas oil into lighter valuable gasoline base components. The fractionation plant includes six process units and 15 streams. Measurements of total flow rates are to be reconciled.

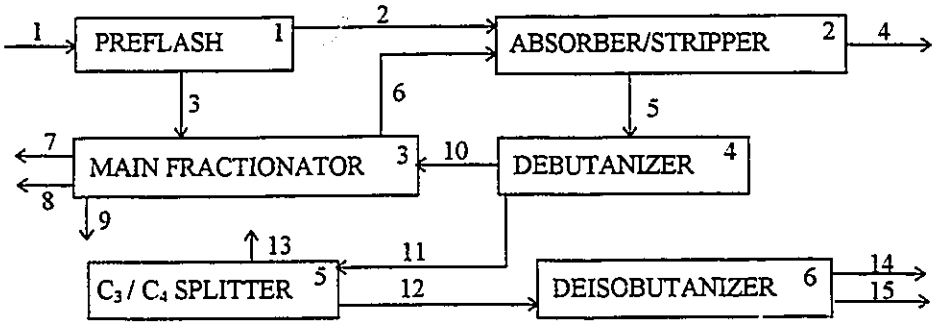


Figure 5.2 Sunoco Hydrocracker Fractionation Plant

5.2.1 Case 1

Table 5.8 shows the flow rate measurements taken at 7:00 a.m., February 26, 1986. It also includes the variances, adjusted measurements and statistics. There are no covariances in this problem. The threshold for a uninormal variate with 6 degrees of freedom is 2.631. The outliers and major contributors are shaded in the tables.

Table 5.9 shows what would happen if the suspect measurements flagged by the different statistics were deleted one at a time.

Table 5.8 Hydrocracker Fractionation Plant: Case 1

	$\tilde{x}$	Variance	$\hat{x}$	$z_a = z_a^*$	y_a	Contrib. to $y_{a,5}$	Contrib. to $y_{a,6}$
1	4769.20	56863	4887.89	0.559	-0.266	0.004	0.009
2	356.02	316.9	424.42	**5.346**	0.750	**-1.759**	**3.601**
3	4391.4	48211	4463.47	0.378	-1.154	0.004	-0.004
4	62.06	9.63	59.96	**-5.386**	0.397	-0.002	0.003
5	634.31	1005.9	602.34	-1.077	**-3.892**	-0.051	-0.213
6	213.47	113.9	237.88	**5.303**	**3.610**	-0.226	0.462
7	394.05	388.2	392.66	-0.782		0.000	0.000
8	2191.2	12004	2148.2	-0.782		0.001	0.001
9	1909.40	9114.4	1876.75	-0.782		0.001	0.001
10	212.67	113.1	192.02	**-4.144**		-0.186	0.403
11	423.73	448.9	410.31	-0.697		-0.229	-0.208
12	374.51	350.6	341.58	-1.984		**-1.543**	-0.723
13	70.69	12.5	68.74	**-2.883**		-0.001	0.005
14	218.92	119.8	211.46	-1.083		0.082	0.241
15	132.87	44.1	130.12	-1.083		0.011	0.033

Table 5.9 Deletion of a Variable, Case 1

Variable deleted	Result	Comment
$\tilde{x}_2$	all tests passed	$\chi^2_m = 1.73$, lowest
$\tilde{x}_{12}$	*some outliers*	
$\tilde{x}_4$	$\hat{x}_4$ *negative*	*infeasible*
$\tilde{x}_6$	all tests passed	$\chi^2_m = 2.19$, higher
$\tilde{x}_{10}$	*some outliers*	
$\tilde{x}_{13}$	*some outliers*	

Table 5.10 Tests of a, Case 1

Statistic	Gross errors identified	Comment
z_a, z_a^*	$\tilde{x}_4$, $\tilde{x}_2$, $\tilde{x}_6$, $\tilde{x}_{10}$, $\tilde{x}_{13}$	Poor. Too many outliers because of the confounding. The primary suspect $\tilde{x}_4$ is wrong
$y_{a,5}$	$\tilde{x}_2$, $\tilde{x}_{12}$	Less confounding than the non-PC tests. The primary suspect $\tilde{x}_2$ is correct
$y_{a,6}$	$\tilde{x}_2$	
Q_a	$\tilde{x}_2$, $\tilde{x}_{12}$	

Table 5.10 compares the statistical tests of the adjustments. The PC tests give fewer outliers and are less confounding. The comments are given based on the outcome of the trials when each and every outlier is deleted.

PC tests of a not only give fewer outliers, but also result in the correct identification of the primary suspect $\tilde{x}_2$. The primary suspect $\tilde{x}_4$ given by $z_{a,4}$ was wrong. If deleted, the reconciliation would be infeasible since $\hat{x}_4$ became negative. Another suspect $\tilde{x}_6$ was given by $z_{a,6}$. If it were deleted, there would be no significant statistics. However, the chi-square value, though not exceeding the threshold, would be higher than when $\tilde{x}_2$ was deleted.

This is a convincing example that the PC tests are sharper in detection and less confounding in identification. Table 5.11 summarizes the results that all statistical tests were passed when the suspect measurement $\tilde{x}_2$ was deleted.

Table 5.11 $\tilde{x}_2$ Deleted, Case 1

	$\tilde{x}$	Variance	$\hat{x}$	$z_a = z_a^*$	y_a
1	4769.20	56863	4914.17	0.683	-0.451
2	—	—	(488.43)	—	0.649
3	4391.4	48211	4425.75	0.180	0.219
4	62.06	9.63	62.04	-0.683	0.608
5	634.31	1005.9	639.78	0.190	0.829
6	213.47	113.9	213.39	-0.180	
7	394.05	388.2	392.78	-0.712	
8	2191.2	12004	2152.04	-0.712	
9	1909.40	9114.4	1879.67	-0.712	
10	212.67	113.1	212.14	-0.164	
11	423.73	448.9	427.64	0.206	
12	374.51	350.6	357.16	-1.062	
13	70.69	12.5	70.48	-0.352	
14	218.92	119.8	222.85	0.599	
15	132.87	44.1	134.32	0.599	

5.2.2 Case 2

The sharpness of the PC test may be shown from a different perspective. We look at the measurements, taken on 7:00 a.m., March 3, 1987, shown in Table 5.12.

Table 5.12 Hydrocracker Fractionation Plant: Case 2

	$\tilde{x}$	Variance	$\hat{x}$	$z_a = z_a^*$	y_a	Contribution to $y_{a,5}$
1	3417.60	29199.0	3536.60	0.78	0.49	0.01
2	445.77	496.80	513.29	**3.96**	0.98	**-1.81**
3	2958.60	21883.0	3023.31	0.51	0.40	0.01
4	40.09	4.02	39.53	**-4.05**	0.10	0.00
5	719.48	1294.10	709.97	-0.28	**-3.42**	0.00
6	220.11	121.10	236.21	**3.85**	-2.28	-0.11
7	197.55	97.60	196.86	-1.07		0.00
8	1395.80	4870.80	1361.54	-1.07		0.00
9	1491.20	5559.00	1452.10	-1.07		0.00
10	241.76	146.10	223.41	**-3.31**		-0.14
11	497.44	618.60	486.56	-0.48		-0.15
12	451.30	509.20	415.53	-1.78		**-1.27**
13	72.55	13.20	71.03	-2.48		0.00
14	215.10	115.70	209.92	-0.92		0.02
15	210.58	110.90	205.61	-0.92		0.02

The results are similar to those of case 1. No statistics will be significant if $\tilde{x}_2$ is deleted. However, if we misdeleted $\tilde{x}_{10}$, as suggested by $z_{a,10}$, only the Q_e statistic would be significant (Q_e = **11835.31** compared to its threshold 10284.89 with $k_e = 2$). Without the PC test, one might risk deleting a good measurement while keeping a corrupted one.

5.3. Ammonia synthesis loop

The ammonia synthesis system considered by Crowe *et al.* (1983) and Crowe (1988) is shown in Figure 5.3. There are four units, seven streams, and four components in the process. Eight of the component flows, $N_2^{(1)}$, $H_2^{(1)}$, $Ar^{(1)}$, $N_2^{(2)}$, $Ar^{(2)}$, $N_2^{(3)}$,

$NH_3^{(4)}$, and $H_2^{(5)}$, were measured, where $X^{(j)}$ denotes the flow rate of the component X in stream j. The purge split ratio (α) was fixed at 0.01992 in our computation, in agreement with Crowe *et al.* The measurements were generated from the true values by the addition of normally distributed random noise with the same variance-covariance structure as in Crowe *et al.* 20% gross error in $NH_3^{(4)}$ and 10% in $N_2^{(1)}$ were added to the corresponding measurements.

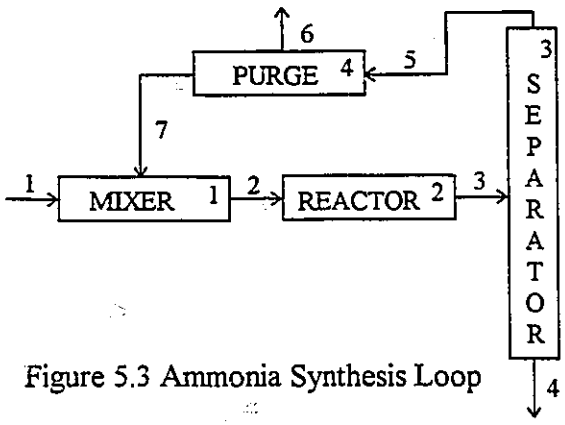


Figure 5.3 Ammonia Synthesis Loop

5.3.1 Case 1: Original Data Used in Crowe et al.

The measurements and the adjusted ones are given in Table 5.13.

Table 5.13 Measured and Reconciled Component Flow Rates

	$N_2^{(1)}$	$H_2^{(1)}$	$Ar^{(1)}$	$N_2^{(2)}$	$Ar^{(2)}$	$N_2^{(3)}$	$NH_3^{(4)}$	$H_2^{(5)}$
$\tilde{x}$	36.69	101.54	0.40	104.76	20.87	76.10	76.04	212.86
$\hat{x}$	36.90	110.51	0.42	109.21	21.04	73.78	70.86	211.75

The incidence and the variance-covariance matrices are given by

$$A = \begin{bmatrix} 1 & -1 & . & . & . & . & 1 \\ . & 1 & -1 & . & . & . & . \\ . & . & 1 & -1 & -1 & . & . \\ . & . & . & . & 1 & -1 & -1 \\ . & . & . & . & \alpha & 1 & . \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 0.82 & 1.14 & 5.12\times10^{-3} & . & . & . & . & . \\ & 6.34 & 1.42\times10^{-2} & . & . & . & . & . \\ & & 1.28\times10^{-4} & . & . & . & . & . \\ & & & 8.16 & 0.816 & . & . & . \\ & & & & 0.326 & . & . & . \\ & & & & & 3.81 & . & . \\ & sym. & & & & & 3.08 & . \\ & & & & & & & 32.0 \end{bmatrix}$$

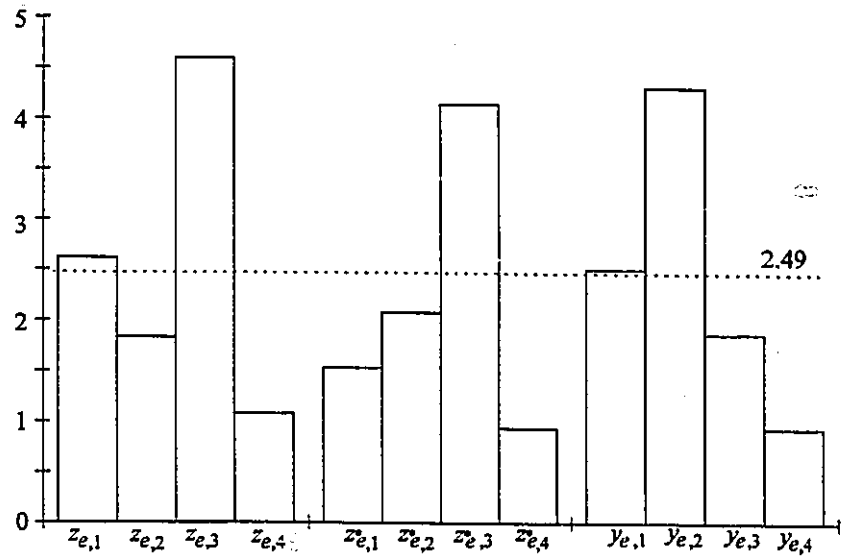


Figure 5.4 The Univariate, MP, and PC Tests of the Reduced Constraints, e, NH_3 Loop

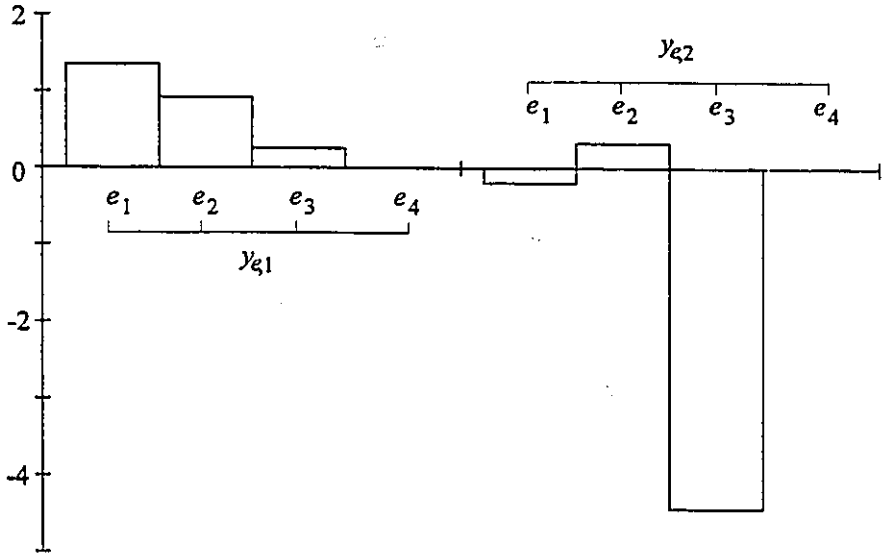


Figure 5.5 Contributions from the Residuals of the Reduced Constraints to y_e, NH_3 Loop

We look at the statistics on e first that are illustrated in Figure 5.4 and Figure 5.5. The four reduced constraints correspond to the mole balances of NH_3, N_2, H_2, and Ar, respectively. While z_e in Figure 5.4 points out that the gross errors are related to the measurements involving NH_3 and H_2, Figure 5.5 shows that $y_{e,1}$ picked up NH_3 and N_2, $y_{e,2}$ picked up H_2. z_e^* in Figure 5.4 picked up nothing except H_2. In fact, the measurements of H_2 were not in gross error but only confounded with that of NH_3. This is an example showing the MP test loses its maximum power when there is more than one gross error in the process. In this case, it not only loses the maximum power, but also is the worst statistic. We notice that z_e failed to pick up N_2. As to the collective tests at 95% confidence level, the truncated chi-square test, $\chi^2_{k_e=2} = \mathbf{24.94}\,(5.99)$, is inflated along with the conventional chi-square test, $\chi^2_{m=4} = \mathbf{29.30}\,(9.49)$. The first two principal components contributed to $\chi^2_{k_e}$, which leads to the same conclusion as the PC test did.

Table 5.14 The Correspondence among the Balances and the Process Units

Equation #	1	2	3	4	5	6
Balance on	$N_2^{(I)}$	$H_2^{(I)}$	$Ar^{(I)}$	$N_2^{(II)}$	$H_2^{(II)}$	$NH_3^{(II)}$
Equation #	7	8	9	10	11	12
Balance on	$Ar^{(II)}$	$N_2^{(III)}$	$H_2^{(III)}$	$NH_3^{(III)}$	$Ar^{(III)}$	$N_2^{(IV)}$
Equation #	13	14	15	16	17	-
Balance on	$H_2^{(IV)}$	$Ar^{(IV)}$	$N_2^{(s)}$	$H_2^{(s)}$	$Ar^{(s)}$	-

We now look at the tests of r. The 17 original constraints correspond to the component mole balances in all the units of the process, as shown in Table 5.14, where $X^{(I)}$ stands for the balance of the component X around the unit I, and $X^{(S)}$ stands for the balance on the purge splitter model, and so on.

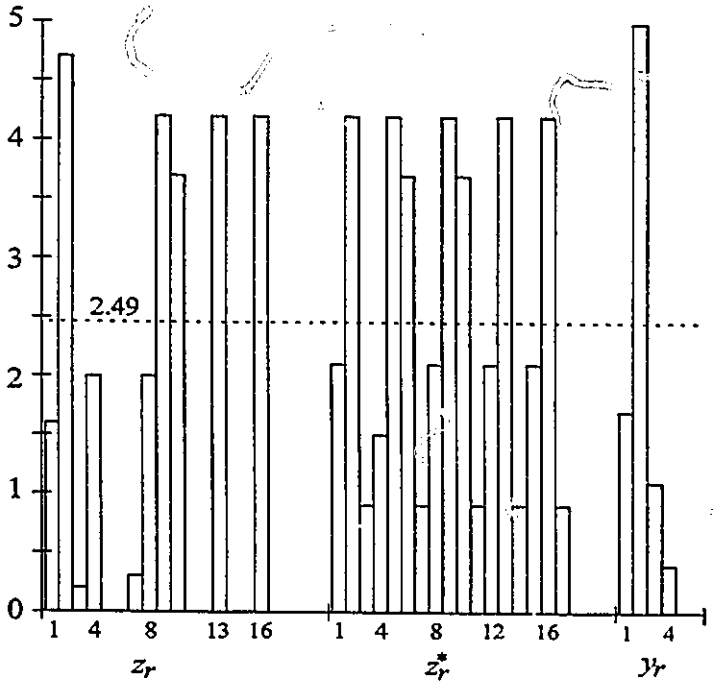


Figure 5.6 The Univariate, MP, and PC Tests of the Original Constraints, r, NH_3 Loop

The z_r, z_r^*, and y_r tests are shown in Figure 5.6. Some of the univariate statistics are not available because H_r is singular. Though Figure 5.6 indicates that z_r, z_r^* and y_r all picked up H_2 and NH_3 (see also Table 5.14), the number of the outliers from the first two tests is large, mainly due to the confounding among the measured variables caused by the process topology. This complicates any further analysis. The maximum number of the principal components that could be retained in this problem is 4. The graph of the contributions to the inflated $y_{r,2}$, shown in Figure 5.7, gives a simpler picture. The two major contributors are the residuals of H_2 in the mixer, and of NH_3 in the separator, which suggests that either the corresponding measurements from these units are in gross error, or the corresponding atoms in the process are not balanced because of gross errors elsewhere.

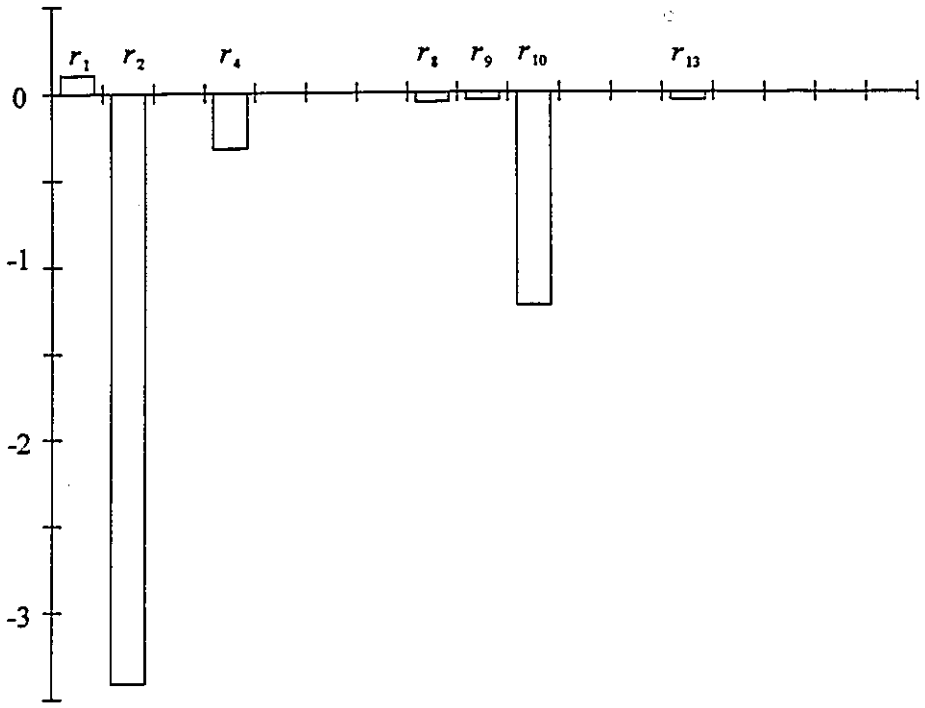


Figure 5.7 Contributions from the Residuals of the Original Constraints to $y_{r,2}$, NH_3 loop

The collective test statistic $Q_{r,k_r=1} = \mathbf{122.47}\,(21.15)$, at an overall 95% confidence level, is inflated, and the contribution graph shown in Figure 5.8 leads to the same conclusion.

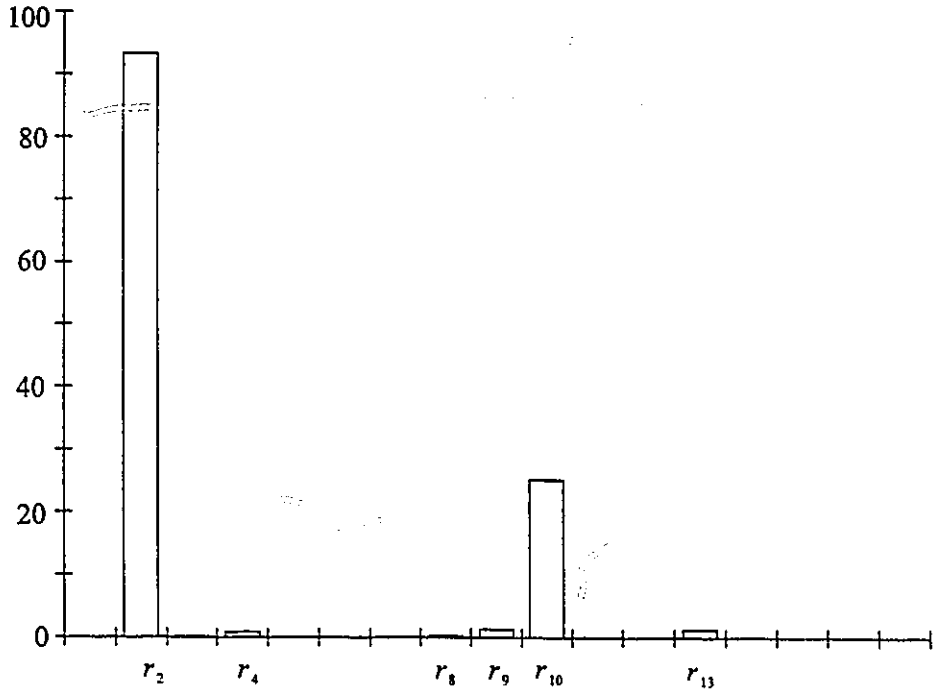


Figure 5.8 Contributions from the Residuals of the Original Constraints to Q_r, NH_3 Loop

Finally we look at the tests of a. z_a and z_a^* shown in Figure 5.9 picked up $H_2^{(1)}$, $H_2^{(5)}$, and $NH_3^{(4)}$. Figure 5.10 shows that the inflated $y_{a,2}$ picked up $H_2^{(1)}$, $NH_3^{(4)}$, and $N_2^{(2)}$. The collective test statistic $\chi^2_{k_a=2} = \mathbf{27.06}\,(5.99)$, at an overall 95% confidence level, is inflated, and the corresponding contribution graph shown in Figure 5.10 indicates that $NH_3^{(4)}$, $H_2^{(1)}$, and $N_2^{(2)}$ are the major contributors. $H_2^{(1)}$ was picked up due to its confounding with $NH_3^{(4)}$, and $N_2^{(2)}$ was picked up since it is correlated with $N_2^{(1)}$ by the mixer.

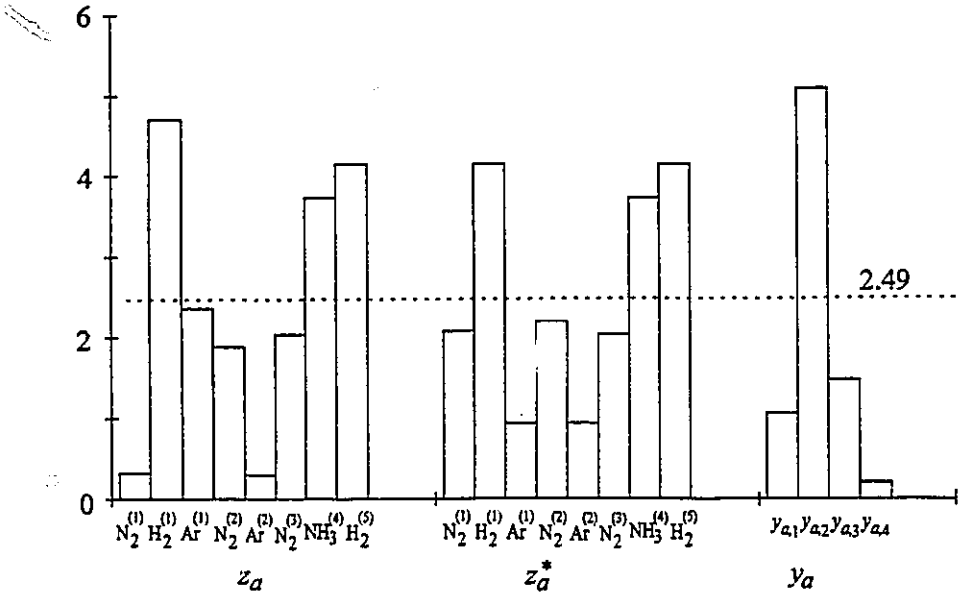


Figure 5.9 The Univariate, MP, and PC Tests of the Measurement Adjustments, a, NH_3 Loop

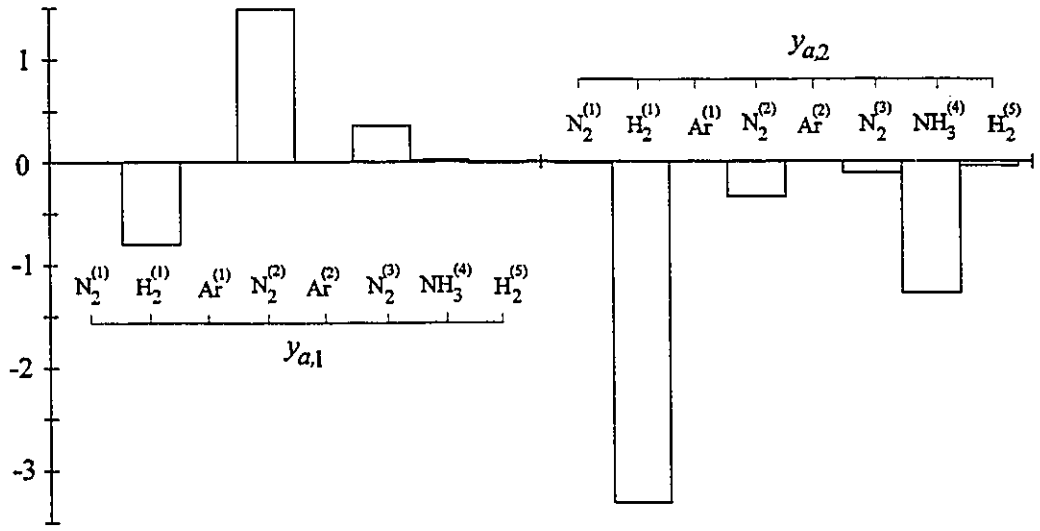


Figure 5.10 Contributions from the Measurement Adjustments to $y_{a,1}$ and $y_{a,2}$, NH_3 Loop

The rank of the covariance matrices is 4. It is obvious that Q_1 and H_r are singular for they do not have full rank. On the other hand, H_e is nonsingular since it does have full rank.

Once we know the measurements are contaminated by gross errors, we have to identify the measurements that are in error, and remove them before data reconciliation is done. Crowe (1988) presented fifteen different deletions of the single and the paired measurements in order to identify the measurements that were in gross error. He concluded that no single deletion reduced the objective function value (χ^2_m) enough but that three different deleted pairs did lead to a sufficient reduction and were thus marked as suspect. The suspect pairs and the corresponding statistics can be found in the paper.

We shall see that the PC tests can do a better job in further distinguishing those three suspect pairs, and correctly identify the one really in error.

It was shown in Crowe (1988) that the deletion of either the prime suspect $\left\{N_2^{(1)}, NH_3^{(4)}\right\}$ or the secondary suspect $\left\{H_2^{(1)}, N_2^{(3)}\right\}$ passed the univariate and the MP tests of the reduced constraints and of the measurement adjustments at 95% confidence level. The conventional chi-square test was passed too. The deletion of the secondary suspect led to a higher chi-square value, 4.84, than that of the prime one, 3.63, though both of them are still smaller than their threshold of 5.99. Table 5.15 summarizes the results of all the collective tests (chi-square, truncated chi-square, and Q tests) of $\boldsymbol{e}$ and $\boldsymbol{a}$ when one suspect pair is deleted. The entries consisting of the word "passed" mean that none of the three collective tests was significant, while the entries consisting of statistics mean that only the specified collective tests were significant, under the threshold given in

the table. It is notable that the PC collective statistics are able to distinguish the three pairs, and correctly identify the pair $\left\{N_2^{(1)}, NH_3^{(4)}\right\}$, which is really in gross error.

Table 5.15 Collective Tests of the Suspect Pairs

(The threshold value $\chi^2_{df=1} = 3.84$; $\alpha = 5\%$)

Suspect Pairs	Test of *e*	Test of *a*
Suspect I: $\left\{N_2^{(1)}, NH_3^{(4)}\right\}$	all tests passed	all tests passed
Suspect II: $\left\{H_2^{(1)}, N_2^{(3)}\right\}$	$\chi^2_{k_e=1} = 4.54$	all tests passed
Suspect III: $\left\{H_2^{(1)}, N_2^{(2)}\right\}$	$\chi^2_{k_e=1} = 5.11$	$\chi^2_{k_a=1} = 4.57$

5.3.2 Case 2: A Subtle Gross Error

To further demonstrate the advantage of using the PC tests, only 7.1% gross errors were added to $N_2^{(1)}$ and $NH_3^{(4)}$ measurements in another study, and this made the gross errors very hard to detect. Among all the statistics only $Q_{a,k_a=2} = \mathbf{4.98}\,(4.85)$, at the overall 95% confidence level, detected them. The contributions from the measurement adjustments to the inflated Q_a are plotted in Figure 5.11. The major contributors to Q_a are $NH_3^{(4)}$, $H_2^{(1)}$, and $N_2^{(1)}$, where $NH_3^{(4)}$ and $N_2^{(1)}$ are indeed in gross error, while $H_2^{(1)}$ was picked up because of its confounding with $NH_3^{(4)}$. Without using the PC test, the identity of the gross errors at such a low level could not be discovered.

If we delete the $NH_3^{(4)}$ measurement, which is the primary contributor to the inflated Q_a statistic, some other test statistics would exceed their thresholds at the overall

95% confidence level, as shown in Table 5.16. We noted that no univariate and maximum power test statistics were significant.

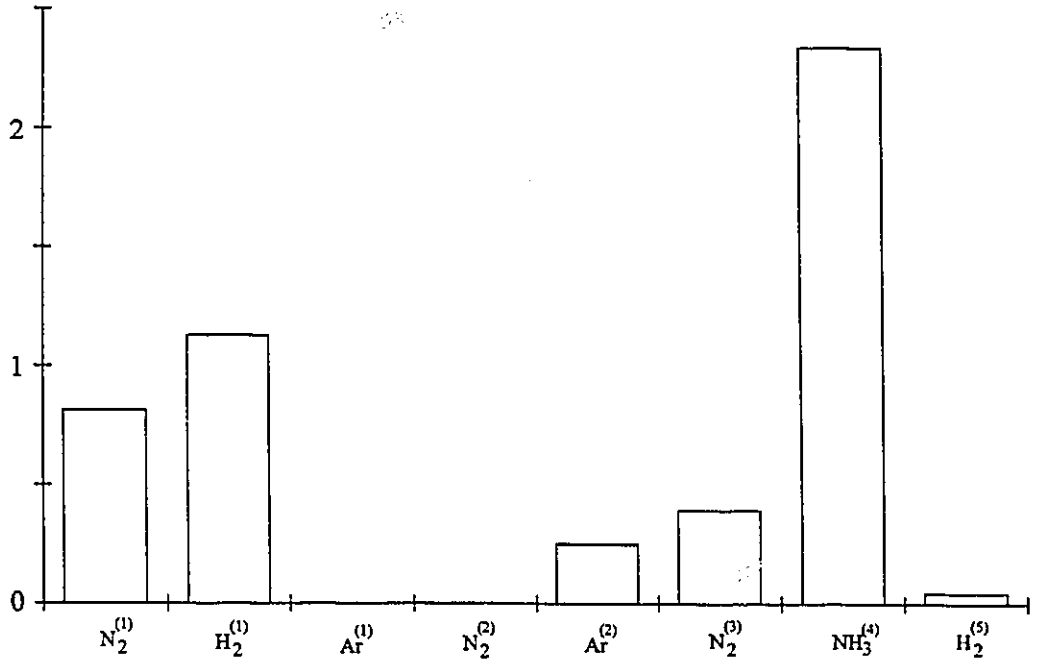


Figure 5.11 Contributions from the Measurement Adjustments to Q_a, NH_3 Loop (7.1% Gross Error Level)

Table 5.16 **Inflated Test Statistics When $NH_3^{(4)}$ is Deleted ($k_r = k_a = 2$)**

Statistic	χ^2_{kr}	χ^2_{ka}	$\chi^2_{m=3}$	$y_{r,2}$	$y_{a,2}$
Value	**9.01**	**8.96**	**9.04**	**2.68**	**2.84**
Threshold	5.99	5.99	7.82	2.39	2.39

The reconciliation after deletion of the suspect pairs { $NH_3^{(4)}$, $N_2^{(1)}$ } and { $NH_3^{(4)}$, $H_2^{(1)}$ } leads to no significance of any test statistic. As a matter of fact, since the

magnitudes of the gross errors are so small, no further distinction between the two pairs could be made based on the measurements at hand.

5.4. Flotation Process A

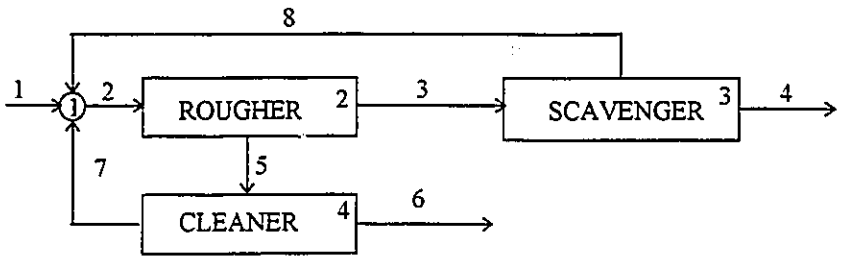


Figure 5.12 Flotation Process A

A flotation circuit having the configuration shown in Figure 5.12 was studied by Smith and Ichiyen (1973). In the process, three circuit sections, rougher, scavenger and cleaner, and three components, copper, zinc and iron, were considered. Only concentration measurements were available for the eight process streams. We will assume, without loss of generality, that the total flow rate of stream 1 is unity. To implement this we will label the total flow rate as a category 1 variable with a small variance so that the value of unity will not be altered within the significant digits at the final solution.

An easy and effective way to deal with the total flow rates is to regard the total flow as a pseudo-component in the process. We will choose $C = 4$ instead of 3 to take account of the total flow rates in this example. Furthermore, we will label the "pseudo-component concentrations" of the total flow rates in stream 2 through stream 8 as category 2 variables with small variances so that their values of unity will not be altered within the significant digits at the final solution. In other word, total flow rates, $\boldsymbol{n}$, are

obtained from Eq. (2-55), while the pseudo-component concentrations for them in δ are kept unity throughout the iterations.

The measured and adjusted variables, as well as the univariate and MP measurement statistics are displayed in Table 5.17. The covariance matrices Σ_1 and Σ_d are diagonal. The variances are proportional to the measurements with the proportional coefficient 6.56%.

Table 5.17 Flotation Process A, Data, Adjustments, and Some Statistics

Species in Process	Stream							
	1	2	3	4	5	6	7	8
copper (m)	0.163	0.447	0.466	0.140	0.960	0.657	1.020	0.440
copper (a)	0.170	0.462	0.406	0.136	0.964	0.649	1.009	0.473
$z_{Cu}=z^*_{Cu}$	0.860	0.665	-2.626	-0.711	0.089	-0.892	-0.242	1.983
zinc (m)	3.93	11.63	8.11	0.49	34.64	52.07	42.73	11.86
zinc (a)	3.856	11.726	8.718	0.491	38.852	51.703	37.011	10.744
$z_{Zn}=z^*_{Zn}$	-1.506	0.300	**3.033**	0.847	**2.853**	-0.586	**-2.775**	-2.700
iron (m)	11.57	12.79	14.35	13.09	14.37	14.67	14.66	13.75
iron (a)	12.335	13.557	13.455	12.179	14.482	14.562	14.470	13.769
$z_{Fe}=z^*_{Fe}$	1.459	1.080	-1.117	-1.454	0.159	-1.014	-0.297	0.029
flow (m)	1	n/m	n/m	n/m	n/m	n/m	n/m	n/m
flow (est.)	1	5.253	4.729	0.934	0.524	0.0657	0.459	3.794

(m: measured; a: adjusted; n/m: not measured; est.: estimated)

There are 16 original constraints and 7 unknown total flow rates, which makes $rank(\boldsymbol{H}_r)=9$. The threshold for a univariate with 95% overall confidence level and 9 degrees of freedom is 2.766.

Table 5.18 Flotation Process A, PC Tests of q

PC	1	2	3	4	5	6	7	8	9
y_q	**2.796**	0.245	-0.979	0.963	-1.469	-1.184	1.053	-2.557	-0.269

The contribution analysis shows that zinc in streams 7 and 5 are the major contributors to $y_{q,1}$ which exceeds its threshold of 2.766.

The truncated chi-square test of q, $\chi^2_{k_q} = \mathbf{9.7624}$, is larger than the table value of 9.49. The contribution analysis shows that the first, third and fourth PC are the major contributors, which correspond to zinc in streams 5 and 7, iron in streams 2 and 3.

The 16 original constraints correspond to Cu, Zn, Fe, and total flow rates in streams 1 to 8, respectively, as shown in Table 5.19. A PC test, $y_{r,4} = \mathbf{-2.86}$, exceeding its threshold, picked up the original constraints 6 and 10. These are the zinc balances around Rougher and Scavenger, respectively. The Q statistic for the original constraints, $Q_r = \mathbf{150.76}$, compared to the 95% confidence value of 72.51. The contribution analysis shows that the major contributors are the balances 2 and 10. They are the zinc balances around Mixer and Scavenger, respectively.

Table 5.19 Flotation Process A, Equation Numbering

	Mixer	Rougher	Scavenger	Cleaner
Cu	1	5	9	13
Zn	2	6	10	14
Fe	3	7	11	15
Total Flow	4	8	12	16

Smith and Ichiyen (1973) acknowledged that the analysis for zinc was the least reliable of the three components. The principal component tests confirmed this statement. In addition, as we mentioned earlier, there is weak evidence from $\chi^2_{k_q}$ that the iron measurements in streams 2 and 3 might also be in error. However, this cannot be positively confirmed based on the information we have.

5.5. Flotation Process B

Smith and Ichiyen (1973) studied another flotation circuit where a metallurgical balance was considered over an entire plant. The flowsheet is shown in Figure 5.13.

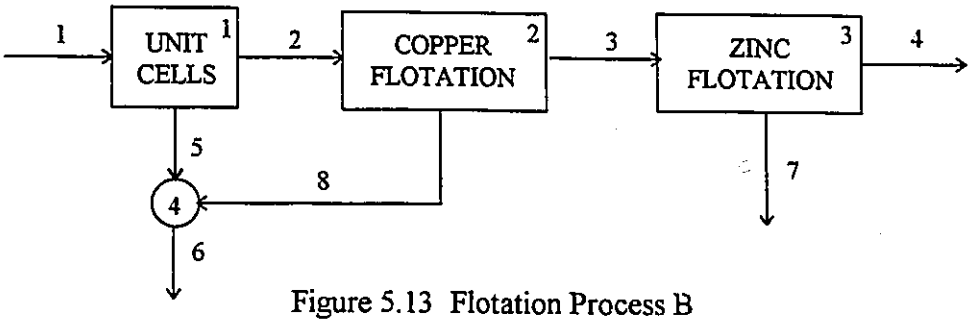


Figure 5.13 Flotation Process B

There are 8 streams, 4 units, and 3 components (copper, zinc, and a pseudo-component for the total flow rates). The measurements were obtained based on X-ray data. Again, the covariance matrices Σ_1 and Σ_d are diagonal. The variances are proportional to the measurements with the proportional coefficient of 6.56%. Both the measured and adjusted variables as well as the univariate and MP measurement statistics are given in Table 5.20.

Table 5.20 Flotation Process B, Data, Adjustments, and Statistics

Species in Process	Stream							
	1	2	3	4	5	6	7	8
copper (m)	1.930	0.450	0.130	0.090	19.860	21.440	0.510	n/m
copper (a)	1.917	0.451	0.126	0.092	19.950	21.458	0.515	0.301
$z_{Cu}=z^*_{Cu}$	-2.282	1.909	-0.579	0.580	1.907	**2.710**	0.580	—
zinc (m)	3.810	4.920	5.360	0.410	7.090	4.950	52.100	n/m
zinc (a)	4.603	4.408	4.578	0.410	7.004	4.885	51.681	**-0.117**
$z_{Zn}=z^*_{Zn}$	**4.203**	-1.909	**-2.593**	-0.580	-1.907	**-2.710**	-0.580	—
flow (m)	1	n/m	n/m	n/m	n/m	n/m	n/m	n/m
flow (est.)	1	0.925	0.916	0.842	0.075	0.084	0.075	0.009

(m: measured; a: adjusted; n/m: not measured; est.: estimated)

There are 12 original constraints, 7 unknown total flow rates, and 2 unknown concentrations, which makes $rank(H_r) = 3$. The threshold for a uninormal variate with 95% overall confidence level and 3 degrees of freedom is 2.388.

Table 5.21 Flotation Process B, PC Tests of q

PC	1	2	3
y_q	-0.545	-1.188	4.034

The contribution analysis indicates that zinc in streams 1, 2 and 3 are the major contributors to $y_{q,3}$ which exceeds its threshold of 2.388.

The Q test of q is **1.379**, which is larger than the threshold of 0.318. The conclusion drawn from the contribution analysis is the same as that from y_q.

Table 5.22 Flotation Process B, Equation Numbering

	Unit Cells	Copper Flotation	Zinc Flotation	Mixer
Cu	1	4	7	10
Zn	2	5	8	11
Total Flow	3	6	9	12

The 12 original constraints correspond to Cu, Zn, and total flow rates in streams 1 to 8, respectively, as shown in Table 5.22. A PC statistic, $y_{r,2} = \mathbf{4.09}$, exceeding its threshold, picked up the original constraints 2 and 8. These are the zinc balances around the unit cells and zinc flotation units, respectively. The truncated chi-square test, $\chi^2_{k_r} = \mathbf{16.86}$, is inflated compared to the 95% confidence value of 5.99. The contribution analysis shows that the second PC is the major contributor, which leads to the same argument. The PC tests of the constraints are much less confounding than the MP and univariate tests, as compared to Table 5.23.

Table 5.23 Flotation Process B, Tests of r

i	1	2	3	4	5	6	7	8	9	10	11	12
z_r	2.13	**-3.26**	**3.66**	-0.67	-0.56	0.64	0.58	**2.40**	-1.90	1.39	-0.22	0.82
z_r^*	2.28	**-4.20**	**4.22**	**2.71**	**-2.71**	**2.71**	0.58	-0.58	-0.58	**2.71**	**-2.71**	**2.71**

As indicated by Smith and Ichiyen (1973), the analysis for zinc was the least reliable of the two components. The result of the principal component tests agrees with their observation. We noticed that although the PC tests and other tests all detected that the zinc measurements may be in error, they do not refer to the same set of the measurements. In addition, based on the PC tests, we would not suspect any copper measurement. Unfortunately, we do not have any further information to compare the tests.

We note that the adjustment to the zinc measurement in stream 8 given in Table 5.20 is infeasible, because it makes the total flow rate negative. This is caused by the poor zinc data.

5.6. Mineral Processing Plant

Hodouin and Everell (1980) studied a mineral process, shown in Figure 5.14, which concentrates copper and lead sulfides of a fine-grained pyritic Zn-Pb-Cu ore. After the roughing stage, there are three cleaning stages and two scavenging stages. The tailings of the cleaner scavenger are combined with those of another parallel circuit, thickened and reground before being recycled as rougher scavenger feed. The pulp flowing in this circuit has been sampled and analysed for Cu, Pb, Zn and seven other chemical elements. A total of 150 assays is thus available around this circuit but no total flow rate measurements were performed. The objective was to optimally use the 150 assays in order to obtain a good measure of the flows in the flotation process. For an alternative way to reconcile the data, see Schraa (1995).

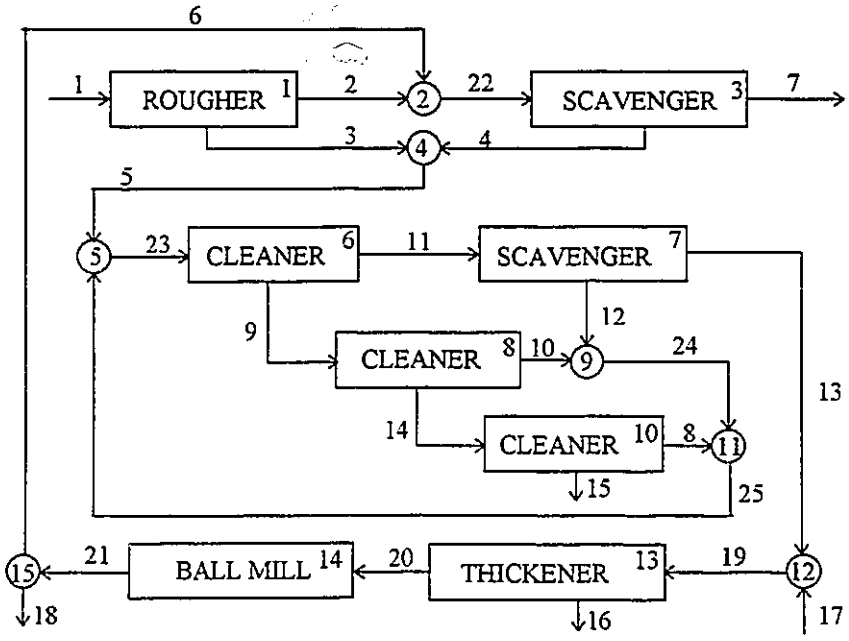


Figure 5.14 Mineral Processing Plant

This is a problem of larger size. There are 15 units, 25 streams (out of which 15 were measured), and 10 chemical components. An extra pseudo-component is needed for the total flow rate. The incidence matrix A is 15×25. The matrix B_1 is 165×11, B_2 165×154, and B_3 165×110. Y^T is 77×165. The reduced balance matrix $Y^T B_1$ is then 77×11, and $Y^T B_2$ is 77×154. There are 11 category 1, 154 category 2, and 110 category 3 variables. H_e is of order of 63, H_r is of 165, all with the rank of 63. Q_1 is of order of 11 with the rank of 10, Q_2 is of 154 with the rank of 63, Q is of 165 with the rank of 63.

The measured and adjusted variables for the base case are displayed in Table 5.24 and Table 5.25, respectively. Each column gives component flow rates in streams 1 through 15.

Table 5.24 Mineral Processing Problem: Measurements

	1	2	3	4	5	6	7	8	9	10
1	0.018	0.0449	0.0903	0.2871	0.0105	0.050	0.07	0.06	0.098	0.09
2	0.014	0.0199	0.0948	0.2944	0.0059	0.035	0.06	0.03	0.085	0.07
3	0.061	0.2510	0.1032	0.2548	0.0480	0.137	0.07	0.23	0.097	0.06
4	0.073	0.1089	0.1331	0.2986	0.0310	0.114	0.08	0.19	0.122	0.07
5	0.067	0.1688	0.1207	0.3014	0.0393	0.128	0.08	0.20	0.120	0.06
6	0.043	0.0757	0.1087	0.3218	0.0211	0.081	0.07	0.16	0.083	0.07
7	0.013	0.0155	0.0954	0.2758	0.0050	0.030	0.06	0.04	0.078	0.08
8	0.056	0.1729	0.1533	0.2543	0.0393	0.118	0.10	0.20	0.114	0.08
9	0.081	0.2315	0.1420	0.2358	0.0497	0.132	0.10	0.30	0.115	0.07
10	0.065	0.1584	0.1464	0.2656	0.0403	0.130	0.10	0.27	0.120	0.08
11	0.059	0.0949	0.1373	0.2984	0.0303	0.095	0.07	0.14	0.104	0.06
12	0.090	0.1526	0.1640	0.2553	0.0445	0.151	0.12	0.24	0.148	0.05
13	0.049	0.0521	0.1117	0.3394	0.0188	0.083	0.07	0.12	0.083	0.05
14	0.100	0.3487	0.1112	0.1866	0.0628	0.165	0.07	0.31	0.098	0.05
15	0.110	0.3669	0.1042	0.1835	0.0665	0.159	0.06	0.34	0.084	0.06
%	6	3	4	2	2	4	26	11	6	70

The covariance matrices Σ_1 and Σ_d are diagonal. The variances are assumed to be proportional to the measurements. The squared roots of the proportional coefficients are given in the last row of Table 5.24.

By inspecting the flowsheet, we could see that the bottom part of the flowsheet with the thickener and the ball mill were essentially not measured, while most of the streams on the remaining flowsheet were measured. The bottom part of the process involving streams 16 through 21 is not observable. The other unmeasured flows (22 through 25) can be computed from some of the measured flows.

The estimated total flow rates of streams 2 through 14, under the assumption that the feed flow rate (stream 1) is unity, are: $\boldsymbol{n}$ = [0.8899 0.1101 0.1355 0.2456 0.1636 0.9180 0.0099 0.2842 0.1763 0.2507 0.1032 0.1475 0.1080 0.0981].

Table 5.25 Mineral Processing Problem: Reconciled Measurements

	1	2	3	4	5	6	7	8	9	10
1	0.020	0.0451	0.0946	0.2819	0.0105	0.045	0.06	0.06	0.087	0.08
2	0.015	0.0195	0.0940	0.2851	0.0059	0.035	0.06	0.03	0.087	0.08
3	0.062	0.2527	0.0999	0.2559	0.0475	0.131	0.07	0.23	0.088	0.05
4	0.073	0.1113	0.1267	0.3018	0.0307	0.109	0.07	0.18	0.104	0.06
5	0.068	0.1747	0.1147	0.2812	0.0382	0.119	0.07	0.20	0.096	0.06
6	0.046	0.0747	0.1095	0.3188	0.0212	0.079	0.07	0.17	0.082	0.07
7	0.012	0.0157	0.0919	0.2886	0.0050	0.032	0.06	0.04	0.083	0.08
8	0.056	0.1731	0.1533	0.2537	0.0392	0.118	0.10	0.20	0.114	0.08
9	0.078	0.2288	0.1360	0.2365	0.0495	0.141	0.09	0.29	0.113	0.07
10	0.066	0.1592	0.1504	0.2651	0.0404	0.125	0.11	0.27	0.122	0.08
11	0.062	0.0938	0.1374	0.3077	0.0298	0.107	0.08	0.16	0.114	0.06
12	0.087	0.1538	0.1639	0.2525	0.0450	0.139	0.10	0.22	0.140	0.05
13	0.045	0.0518	0.1189	0.3463	0.0192	0.085	0.07	0.11	0.096	0.06
14	0.099	0.3424	0.1125	0.1898	0.0644	0.166	0.07	0.32	0.098	0.06
15	0.103	0.3595	0.1084	0.1833	0.0669	0.171	0.07	0.33	0.096	0.06

Hodouin and Everell (1980) did not provide statistical tests. The global test, $\chi^2_{m=63} = \mathbf{142.74}$, exceeds its threshold of 82.53, strongly indicates the existence of gross errors.

We now compare the principal component test against the univariate and MP tests. The result of z_δ test with 8 outliers is given in Table 5.26[4]. The y_δ test has 4 outliers. The major contributors to them are shown in Table 5.27 through Table 5.30. The numbers in the cells indicate the relative magnitude of each major contributor (1 is the largest, 2 is

[4] In this example, z_δ is identical to z^*_δ.

the second largest, and so on). The threshold of these tests is 3.348 with 63 degrees of freedom.

Table 5.26 Outliers in z_δ (z_δ^*) Test

stream/species	1	2	3	4	5	6	7	8	9	10
1										
2										
3									■	
4									■	
5				■					■	
6										
7										
8										
9										
10										
11						■				
12						■				
13									■	
14										
15									■	

Table 5.27 Major Contributors to $y_{q,28}$

stream/species	1	2	3	4	5	6	7	8	9	10
1										
2										
3									7	
4									1	
5				3					2	
6										
7										
8										
9										
10										
11				6					8	
12										
13				5					4	
14										
15										

We observe that all of the outliers were related only to certain components in certain streams. For example, outliers of z_8 are related to the components 4, 6, and 9 in streams 3, 4, 5, 11, 12, 13 and 15.

Table 5.28 Major Contributors to $y_{q,31}$

stream/species	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4				2						
5				1						
6										
7										
8										
9										
10										
11				3						
12										
13										
14										
15										

Table 5.29 Major Contributors to $y_{q,39}$

stream/species	1	2	3	4	5	6	7	8	9	10
1										
2										
3						5				
4						11				
5		13				3				
6										
7				9						
8										
9		10	8			7				
10										
11				2		6				
12						4				
13				12						
14										
15						1				

In order to identify the gross errors, we tried to delete each of the measurements corresponding to the outliers without success. The reconciliation with single deletion could not pass some PC and non-PC statistical tests. It is obvious that there are multiple gross errors. We then tried to delete all the measurements simultaneously corresponding to the z_δ outliers. The result was similar. Finally, we tried to delete all the measurements simultaneously corresponding to the outlier PCs, except for those related to the components 1 and 2[5]. All the PC tests and the regular chi-square test were passed. However, some non-PC tests (the univariate and MP tests) were failed.

Table 5.30 Major Contributors to $y_{q,59}$

stream/species	1	2	3	4	5	6	7	8	9	10
1						6				
2										
3										
4						4				
5		1				7				
6	5					11				
7										
8										
9										
10										
11										
12										
13	9									
14		10								
15	2	8				3				

Our conjecture is that the PC tests might have correctly identified the gross errors, and the results given by the non-PC tests might be wrong. This is based on the fact that the regular chi-square test was passed and that the non-PC tests have been generally less

[5] If we also include those measurements related to the components 1 and 2 in deletion, the result is similar.

powerful than the PC tests. In addition, the gross errors identified by the PC tests are all related to certain species. This suggests that those species were not measured accurately and caused the problems. It seems again that the PC tests are sharper and less confounding in detecting and identifying the gross errors.

Because we do not have access to the process nor any further data, we are not able to prove the conjecture. Nevertheless, the following case studies may support our claim to some extent.

We slightly modify the flowsheet. The thickener and ball mill units were deleted, because the streams linked to these units were not measured nor observable. The reconciled data shown in Table 5.25 satisfied the material balances and were used as the "true" values of the measured variables. Normally distributed random errors were then added to each measurement. When this base case was reconciled, as one may expect, it passed all the statistical tests.

A single error with its magnitude being less than two standard deviations above its "true" value was then added to the concentration measurement of component 9 in stream 3. We should note that the error was not considered as a gross error because of its size. After reconciliation, all the PC tests and the regular chi-square test were passed. However, some MP tests of the original constraints were failed. It is obvious that the MP tests gave a false alarm. This should not be a surprise if one recalls how inaccurate a MP test could be when constraint residuals are moderately to highly correlated (Figure 4.2). In our example, such correlations can be observed from H_r. Furthermore, the error should cause larger residuals of the material balances for component 9 in units 1 and 4 (black cells in Table 5.31). The outliers given by the MP tests were elsewhere (grey cells in Table 5.31). These outliers given by the MP tests were actually caused by confounding, and would not

help in identifying the true gross errors. Table 5.31 indeed shows how confounding a MP test could be.

Table 5.31 Outliers in z_r^* test

unit /species	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										

Table 5.32 Major Contributors to $y_{q,28}$

stream/species	1	2	3	4	5	6	7	8	9	10
1										
2										
3									6	
4									1	
5									7	
6										
7										
8										
9										
10										
11										
12										
13									2	
14	4								5	
15	3									

To further demonstrate the last point, we increased the error to about three times its standard deviation. This time the error should be viewed as a gross error because of its size. The MP tests still gave the same false identification. On the other hand, a PC test,

$y_{q,28}$, detected the gross error, with three major contributors from component 9 in streams 3, 4 and 5, which were related by unit 4, as shown in Table 5.32. There were 7 major contributors to the outlier PC. Because the gross error was related to component 9 in unit 4, the measured concentrations for that component in streams 3, 4 and 5 were major contributors. The remaining contributors were due to confounding because of the process topology. This demonstrated that the PC test is sharper and less confounding compared to the non-PC tests. However, as one may have observed, confounding cannot be completely eliminated even with the PC tests because of the process topology.

5.7. Summary

Six examples were presented to demonstrate how the theories given in Chapters 2, 3 and 4 could be applied to linear and bilinear data reconciliation of different processes. Examples 1, 2 and 3 were linear data reconciliation problems, while examples 4, 5 and 6 were bilinear data reconciliation problems. In particular, among other things, examples 1 and 3 were used to show how the PC tests could detect and identify subtle gross errors. Example 2 was used to illustrate that the PC tests were less confounding. Example 6 was a large-scale process with poor data and many gross errors. It was shown that only the PC tests had detected and identified all of the gross errors and led to a feasible data reconciliation.

CHAPTER 6

DETECTING PERSISTENT GROSS ERRORS

A transient gross error, caused by process disturbances, occurs randomly for an instant of time and does not persist. A persistent gross error, however, is deterministic and usually relates to sensor problems or process leaks. When a persistent gross error occurs, we should detect it as early as possible in order to limit its influence on process operation and product quality, and prevent it from contaminating databases.

It is important to know if a gross error is transient or persistent, and what are the probabilities of type I and type II errors that we may commit in detecting such an error. The statistical tests given in Chapters 3 and 4 cannot efficiently distinguish a persistent gross error from a transient one, because they test one set of measurements at a time, without taking into account the successive sets. It is impossible to estimate the probability of type II error in the hypothesis testing for the same reason.

In this chapter, we study how sequential analysis and PCA can be combined to detect persistent gross errors as early as possible.

6.1. Sequential Analysis

Sequential analysis was pioneered by Wald (1947) and advanced by other researchers (Jackson and Bradley, 1961a, 1961b; Whittle, 1982,1983; Siegmund, 1985,

Liu and Blostein, 1992). It is a method of statistical inference where the number of observations required is not determined in advance but is dependent on the outcome of the observations as they are made. A key merit of this procedure is that it requires in general many fewer observations than other procedures based on a fixed number of observations. It was reported (Wald, 1947) that a sequential test frequently results in a saving of 50 per cent in the number of observations over the most efficient non-sequential test procedure such as the ones based on Neyman - Pearson theory. Therefore sequential analysis is capable of providing an earlier alarm and greater efficiency in gross error detection.

When the variance of a variable is not known, the sequential procedures have been developed by Wald (1947) and Rushton (1950, 1952).

6.1.1 Hypotheses

We showed in Chapter 4 that any of the principal components that we have considered in this thesis[1] is distributed with the zero mean and the unit variance if measurements are free of gross errors. Furthermore, we showed that a principal component is normally distributed if the measurements are.

It is natural therefore to test whether the mean of a principal component is zero. Let y be a principal component, $y \sim N(\theta, 1)$, with an unknown mean θ and unit variance. In general, the greater the absolute deviation of θ from zero, the greater possibility of having gross errors. We are interested in testing the null hypothesis H_0 that $\theta = 0$ against the alternative hypothesis H_1 that $\theta = \theta_1 \neq 0$. If $\theta \neq 0$ but is near zero, a test is practically indifferent in accepting or rejecting the null hypothesis of no gross error. The acceptance of the null hypothesis would not be a serious error. However, there will be a positive value

[1] We limit ourselves to the principal components derived from $\boldsymbol{e}$, $\boldsymbol{r}$ and $\boldsymbol{a}$ only.

δ such that the acceptance of the null hypothesis is regarded as an error of practical importance whenever

$$|\theta| \geq \delta \tag{6-1}$$

therefore we may test the null hypothesis that $\theta = 0$ against the alternative that $|\theta| \geq \delta$ in practical applications.

6.1.2 A Sequential Sampling Scheme with Prescribed α and β

We need to derive a sampling scheme for which the probability that the null hypothesis will be rejected does not exceed a small prescribed value α whenever $\theta = 0$, and the probability of accepting the null hypothesis does not exceed a small prescribed value β whenever the alternative hypothesis is true[2]. In other words, we are to derive a sampling plan to perform hypothesis testing with the probability of type I error α and the probability of type II error β.

A particular method of sequential analysis, known as the sequential probability ratio test, was developed by Wald (1947). The method can be applied to a principal component for detecting persistent gross errors[3].

6.1.3 Sequential Probability Ratio Test (SPRT)

Let $f(y, \theta)$ denote the probability density function of a random variable y (here y is a principal component, but this is not necessary). Let H_0 be the null hypothesis that

[2] For simplicity, we use α, not α^*, to represent a probability of type I error for a variable in this chapter.

[3] The method of course can be applied directly to $\boldsymbol{e}$, $\boldsymbol{r}$ and $\boldsymbol{a}$. However, as we have seen in Chapter 5, tests for principal components usually result in sharper detection and less confounding identification. In multivariate sequential analysis that we are about to see in this chapter, the use of principal components is the only way to detect gross errors.

$\theta = 0$, and H_1 be the alternative hypothesis that $\theta = \theta_1 \neq 0$. Thus, the probability density function of y is given by $f(y,0)$ when H_0 is true, and by $f(y,\theta_1)$ when H_1 is true. We denote the successive observations on y by $y^1, y^2, \cdots$, etc.

Let $p_{1j} = p_{1j}(y^1, \cdots, y^j)$ denote the joint probability density function that a sample $y^1, \cdots, y^j$ is obtained for any integer $j > 0$ when H_1 is true, and let $p_{0j} = p_{0j}(y^1, \cdots, y^j)$ denote that when H_0 is true. Of course if the successive observations $y^1, y^2, \cdots$, are independent observations on y, we would have $p_{1j} = f(y^1,\theta)\cdots f(y^j,\theta)$ and $p_{0j} = f(y^1,0)\cdots f(y^j,0)$. The sequential probability ratio test for testing H_0 against H_1 is so defined that at each stage of sampling the probability ratio $\frac{p_{1j}}{p_{0j}}$ is computed and compared to two constants A and B ($0 < B < A$), as illustrated in Figure 6.1.

$p_{1j} / p_{0j} \leq B$	$B < p_{1j} / p_{0j} < A$	$p_{1j} / p_{0j} \geq A$
accept H_0 (no gross error)	take one more observation	reject H_0 (gross error)
	B	A

Figure 6.1 Sequential Probability Ratio Test

If $B < \frac{p_{1j}}{p_{0j}} < A$, the sampling is continued by taking an additional observation. If $\frac{p_{1j}}{p_{0j}} \geq A$, the process is terminated with rejection of H_0 (acceptance of H_1). If $\frac{p_{1j}}{p_{0j}} \leq B$, the process is terminated with the acceptance of H_0.

The constants A and B can be so determined that the test will have the strength (α, β). The relations among the quantities α, β, A and B are given by

$$A \leq \frac{1-\beta}{\alpha} \tag{6-2}$$

$$B \geq \frac{\beta}{1-\alpha} \tag{6-3}$$

and for all practical purposes we may choose

$$A = \frac{1-\beta}{\alpha} \tag{6-4}$$

$$B = \frac{\beta}{1-\alpha} \tag{6-5}$$

This results in a slight increase in the number of observations, which would be acceptable for flow rate sampling that is not excessively costly.

It can be shown that if the successive observations $y^1, y^2, \cdots$ are independent observations on y, the sequential process will eventually terminate (Wald, 1947). However, this requires the assumption of independent sampling that is not quite practical. The conditional distribution of the jth observation y^j is in general affected by the outcome of the preceding observations $y^1, \cdots, y^{j-1}$, which makes the successive observations dependent. Even in a stationary state, we cannot rule out dependence in time among variables[4]. Fortunately, the inequalities (6-2) and (6-3) remain valid in spite of the dependence of the successive observations, provided that the probability is one that the procedure will eventually terminate. The method is valid in general for dependent observations.

[4] A variable will be autocorrelated.

6.2. Collecting Data for Sequential Analysis

Our interest in using sequential analysis is to detect quickly and efficiently any persistent gross error. To achieve this goal, some caution must be exercised. If a sampling scheme of high frequency is applied to a slow process, a transient gross error may be reflected in a number of consecutive sets of measurements. This may cause the sequential analysis to flag it wrongly as a persistent gross error. To avoid this problem, the sampling frequency must be much slower than the process. When sampling frequency is high, only data sets separated from one another by an appropriate time interval should be used in sequential analysis. Alternatively, averages of the observations may be used.

6.3. Univariate Sequential Test

6.3.1 Test that the Mean of a Principal Component is Zero

An adequate sampling scheme for testing that the mean of a principal component is zero is given as follows. We compute the ratio

$$\frac{p_{1j}}{p_{0j}} = \frac{1}{2}\frac{e^{-\frac{1}{2}\sum_{i=1}^{J}(y^i-\delta)^2} + e^{-\frac{1}{2}\sum_{i=1}^{J}(y^i+\delta)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{J}(y^i)^2}} \tag{6-6}$$

as each set of measurements[5] is available. The computation will continue as long as $B < \frac{p_{1j}}{p_{0j}} < A$. The null hypothesis that the principal component is not in gross error is

[5] We can only measure flow rates and concentrations, not principal components. However, for simplicity, we will speak in terms of the measurements of principal components. This should not cause any confusion.

accepted if $\frac{p_{1j}}{p_{0j}} \le B$, and rejected if $\frac{p_{1j}}{p_{0j}} \ge A$, where A and B are given by Eqs. (6-4) and (6-5).

The expression for $\frac{p_{1j}}{p_{0j}}$ can be simplified to

$$\frac{p_{1j}}{p_{0j}} = e^{-\frac{j\delta^2}{2}} \cosh\left(\delta \sum_{i=1}^{j} y^i\right) \tag{6-7}$$

Substituting this value of $\frac{p_{1j}}{p_{0j}}$ in the above inequalities and taking logarithms, we find that those inequalities become

$$\log B + j\frac{\delta^2}{2} < \log\left[\cosh\left(\delta \sum_{i=1}^{j} y^i\right)\right] < \log A + j\frac{\delta^2}{2} \tag{6-8}$$

$$\log\left[\cosh\left(\delta \sum_{i=1}^{j} y^i\right)\right] \ge \log A + j\frac{\delta^2}{2} \tag{6-9}$$

$$\log\left[\cosh\left(\delta \sum_{i=1}^{j} y^i\right)\right] \le \log B + j\frac{\delta^2}{2} \tag{6-10}$$

For each set of measurements we compute

$$Z_j = \log\left[\cosh\left(\delta \sum_{i=1}^{j} y^i\right)\right] \tag{6-11}$$

and a conclusion is made the first time that Z_j does not lie between $\log A + j\frac{\delta^2}{2}$ and $\log B + j\frac{\delta^2}{2}$. The hypothesis that the principal component is not in gross error is accepted if $Z_j \le \log B + j\frac{\delta^2}{2}$, and rejected if $Z_j \ge \log A + j\frac{\delta^2}{2}$.

6.3.2 *Hydrocracker Fractionation Plant*

In Chapter 5 we presented the Sunoco Hydrocracker Fractionation Plant studied by Bailey (1991). The analysis of data sampled at 7 a.m., Feb. 26, 1986 showed that two principal components in the measurement tests ($y_{a,5}$ and $y_{a,6}$) were in gross error, and the analysis of data sampled a year later (at 7 a.m., Feb. 27, 1987) showed that only one principal component in the measurement tests ($y_{a,5}$) was in gross error. We want to know if the gross error is persistent or transient. Furthermore, we want to know why there was only one outlier principal component in one case while there were two in the other.

To answer these questions we need some additional samples and to use the sequential method.

6.3.2.1 Choose α, β and δ

The choice of the parameters α, β and δ has a noticeable impact on sequential analysis, just as the choice of α has on the non-sequential tests in Chapters 3 and 4. An inappropriate choice may result in more samples being taken or a wrong inference. We recommend the following guidelines in choosing a consistent set of parameters α, β and δ.

1) α should be obtained from the prescribed probability of the overall type I error, as discussed in Chapter 4. In our example, we specify the desired probability of the overall type I error as 0.05. With 6 degrees of freedom (6 principal components could be retained) we have $\alpha = 0.0085$ with the threshold of 2.631.

2) β should be small enough to keep the test sharp. This is because β is the probability of the error that one could commit if a principal component is subject to persistent gross errors while one infers otherwise. A gross error can significantly distort

data reconciliation and usually relates to sensor failure or a process leak. We choose $\beta = 0.001$ in this example.

3) The choice of δ reflects one's judgment that how far away the absolute value of the mean of a principal component from zero is considered to be evidence of gross error. Since a principal component y is normalized with unit variance, δ should be a small positive number. It would be a reasonable expectation that $0 < \delta \leq \delta_\alpha$, where δ_α is the threshold value for the principal component corresponding to α. In this example, we choose $\delta = 2.6$.

6.3.2.2 Sequential Analysis for the Principal Components

The data obtained at 7 a.m. and 7 p.m. on Feb. 26, 1986 (Day 1), Feb. 27, 1987 (Day 2) and March 3, 1987 (Day 3) are given in Tables 6.1 and 6.2 (Bailey, 1991). The six sets of measurements are labeled A through F.

Table 6.1 Hydrocracker Fractionation Plant Data Sampled in the Mornings

A: 7 a.m., Day 1		B: 7 a.m., Day 2		C: 7 a.m., Day 3	
$\tilde{x}$	σ^2	$\tilde{x}$	σ^2	$\tilde{x}$	σ^2
4769.20	56863.00	3391.00	28747.00	3417.60	29199.00
356.02	316.90	258.71	167.30	445.77	496.80
4391.40	48211.00	3060.40	23415.00	2958.60	21883.00
62.06	9.63	46.33	5.37	40.09	4.02
634.31	1005.90	600.81	902.40	719.48	1294.10
213.47	113.90	290.07	210.40	220.11	121.10
394.05	388.20	247.89	153.60	197.55	97.60
2191.20	12004.00	1481.90	5490.10	1395.80	4870.80
1909.40	9114.40	1458.00	5314.60	1491.20	5559.00
212.67	113.10	278.17	193.40	241.76	146.10
423.73	448.90	336.20	282.60	497.44	618.60
374.51	350.60	289.78	209.90	451.30	509.20
70.69	12.50	58.66	8.60	72.55	13.20
218.92	119.80	107.66	28.98	215.10	115.70
132.87	44.10	165.79	68.72	210.58	110.90

In the first case study, we look at the data sorted in the morning - evening order. This arrangement would reveal any shift to shift difference. The results of sequential analysis are given in Figures 6.2 through 6.7. For the illustration purpose, Figures 6.2, 6.3, 6.4 and 6.5 only show a single round of sequential test - the first time when a sequential test accepts or rejects H_0.

Table 6.2 Hydrocracker Fractionation Plant Data Sampled in the Evenings

D: 7 p.m., Day 1		E: 7 p.m., Day 2		F: 7 p.m., Day 3	
$\tilde{x}$	σ^2	$\tilde{x}$	σ^2	$\tilde{x}$	σ^2
4787.98	57308.91	3402.65	28937.75	3344.21	27954.48
354.99	313.29	261.77	171.61	389.99	380.25
4417.33	48774.17	3063.72	23459.75	2963.41	21958.23
66.67	10.89	44.60	4.84	38.34	3.61
638.93	1018.57	636.63	1014.42	678.03	1151.24
213.41	114.03	288.38	208.85	222.21	123.34
400.98	402.08	251.24	158.27	369.00	339.31
2191.17	12000.76	1511.56	5716.81	1316.84	4339.11
1909.24	9119.52	1458.91	5316.41	1495.39	5591.67
208.49	109.18	313.58	246.76	277.60	192.12
434.82	472.11	336.14	282.24	396.97	392.99
384.48	368.99	286.82	206.32	345.76	298.29
71.54	12.87	63.92	10.06	68.48	11.78
215.18	116.38	102.94	26.65	138.45	47.64
144.11	52.12	169.49	71.57	186.83	87.65

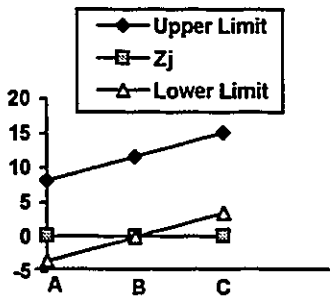


Figure 6.2 Sequential Test for $y_{a,1}$

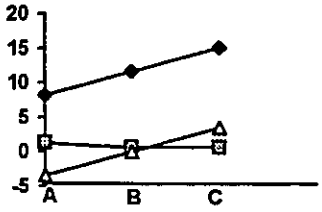


Figure 6.3 Sequential Test for $y_{a,2}$

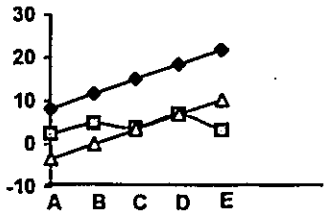


Figure 6.4 Sequential Test for $y_{a,3}$

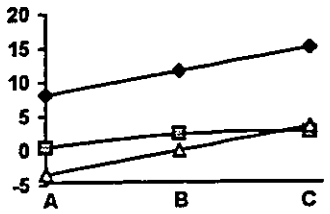


Figure 6.5 Sequential Test for $y_{a,4}$

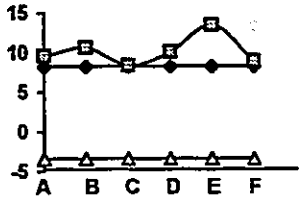


Figure 6.6 Sequential Test for $y_{a,5}$

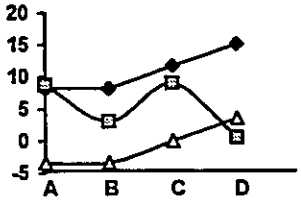


Figure 6.7 Sequential Test for $y_{a,6}$

Principal components 1 through 4 are not subject to gross error, which is the same as what we observed in Chapter 5.

Principal component 5 and 6 are subject to gross error, again similar to what we observed in Chapter 5. If we take a closer look at Figures 6.6 and 6.7, we may find that the statistics for principal component 5 are always above the upper limit, which indicates that the principal component is subject to a persistent gross error. However, the statistics for principal component 6 are above the upper limit in the first round of test (A), but below the lower limit in the other round (B, C and D). The result shows no evidence of shift to shift difference. However, is principal component 6 also subject to a persistent gross error?

We investigate this problem by performing another case study, where the data are arranged chronologically, i.e., in the sequence of A, D, B, E, C, F. The result is given in Figure 6.8. In the first two rounds (A and D) the sequential test statistics corresponding to the data sampled on Day 1 are above the upper limit, while in the next round (B, E, C and F) the sequential test statistics corresponding to the data sampled on Day 2 and Day 3 are in between the upper and lower limits. This suggests that principal component 6 was likely

subject to persistent gross error on Day 1, and might or might not be in persistent gross error on Day 2 and Day 3. As a matter of fact, it would be indifferent based on the information that we have to infer if the principal component was in gross error. Our best guess is that the principal component might be in persistent gross error considering the trend of Z_j's on Figure 6.8 and the fact that the data were sampled in a period of a year. It would be a good idea to check the sensors, and the process units related to that principal component. If possible, more measurements should be taken to get a better inference.

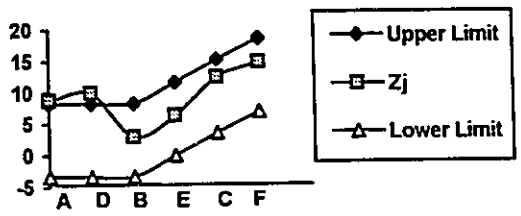


Figure 6.8 Another Sequential Test for $y_{a,6}$

Having detected the existence of gross errors, we need to identify them. This is discussed in 6.5.

6.4. Collective Sequential Test

It is time-consuming to analyze multiple charts simultaneously, especially when a problem is large in size. In Chapters 3 and 4 we investigated a number of non-sequential collective statistical tests. They are used to check whether measurements are in error at a particular instant of time. In this section, we study a collective sequential test. In practical situations, we may switch to the univariate sequential test if the statistic of the collective

sequential test is inflated. This will give us a simpler picture about the process and the measurements, and save computation time. It is particularly useful in on-line monitoring and real-time data reconciliation.

6.4.1 Sequential Chi-Square Test for Principal Components

The sequential chi-square test was developed by Jackson and Bradley (1961a, 1961b). It is also a SPRT. As we know, the univariate sequential test can be used to test a_i (a measurement adjustment) and $y_{a,j}$ (a principal component of the measurement adjustments), but the test for $y_{a,j}$ is sharper and less confounding. On the other hand, the sequential chi-square test can only be used to test y_a, not a. This is because the inverse of the covariance matrix is required in the test. The covariance matrix of y_a is always unity, while the covariance matrix of a is always singular. This is yet another reason why principal components are used in detecting gross errors.

By the same token, the sequential chi-square test can only be used to test y_r, not r, unless there are no unmeasured variables (category 3 variables). Though the covariance matrix of e is nonsingular, the tests for e and y_e are often not as good as the tests for the measurement adjustments and the original residuals, because e corresponds to the reduced constraints only.

6.4.2 Hypotheses

In the univariate sequential test, the null hypothesis is that $H_0: \theta = 0$, and the alternative is that $H_1: \theta \neq 0$ or $|\theta| \geq \delta$. In multivariate case, the single parameter θ is replaced by a vector. The null hypothesis is that $H_0: \theta = \mathbf{0}$. However, the alternative hypothesis that $H_1: |\theta| \geq \delta$ would be very difficult to test. It is easier to operate with the

surface of k_a - dimensional ellipsoids. The null hypothesis that $H_0: \theta = 0$ is identical to $E(y_a)^T E(y_a) = 0$. The alternative hypothesis can be the same form but equal to a larger scalar value: $H_1: E(y_a)^T E(y_a) = \lambda_1^2$ where $\lambda_1 > 0$. The covariance matrix of the principal components does not appear in the expressions because it is identity. If we recall the definition of a chi-square variable, the inverse of the covariance matrix is always required unless it is identity. Therefore, using the principal components is the only way to perform collective sequential analysis on the measurement adjustments.

6.4.3 The SPRT Test

The probability ratio in sequential chi-square analysis can be expressed by (Jackson and Bradley, 1961b)

$$\frac{p_{1j}}{p_{0j}} = {}_0F_1\left(\frac{k_a}{2}, \frac{j\lambda_1^2\chi_j^2}{4}\right) exp\left(\frac{-j\lambda_1^2}{2}\right) \tag{6-12}$$

where

$$\chi_j^2 = j\bar{y}_a^T \bar{y}_a \tag{6-13}$$

$\overline{y_a}$ is a k_a - element vector of sample means based on j observations

$$\bar{y}_a = \frac{1}{j}\sum_{i=1}^{j} y_a^i \tag{6-14}$$

and

$${}_0F_1(c,x) = 1 + \frac{x}{c} + \frac{x^2}{c(c+1)2!} + \cdots + \frac{x^i}{c(c+1)\cdots(c+i-1)i!} + \cdots \tag{6-15}$$

which is known as the generalized hypergeometric function.

It is shown in Jackson and Bradley (1961a) that the test terminates with the probability of unity.

The evaluation of ${}_0F_1(c,x)$ involves the generalized hypergeometric series that converges slowly. A closer look at (6-12) reveals that in the probability ratio only χ_j^2 is unknown. We may obtain the following equations based on (6-4), (6-5), (6-12) and the prescribed α and β, for $j = 1,2,\cdots$, and compute the lower and upper limits $\underline{\chi_j^2}$ and $\overline{\chi_j^2}$ for χ_j^2 in advance

$$ {}_0F_1\left(\frac{k_a}{2}, \frac{j\lambda_1^2\underline{\chi_j^2}}{4}\right) exp\left(\frac{-j\lambda_1^2}{2}\right) = \frac{\beta}{1-\alpha} \tag{6-16} $$

$$ {}_0F_1\left(\frac{k_a}{2}, \frac{j\lambda_1^2\overline{\chi_j^2}}{4}\right) exp\left(\frac{-j\lambda_1^2}{2}\right) = \frac{1-\beta}{\alpha} \tag{6-17} $$

The nonlinear equations in (6-16) and (6-17) are solved in our program by the Gauss-Newton method with a mixed quadratic and cubic line search procedure. A problem associated with solving (6-16) is that some $\underline{\chi_j^2}$ values might be negative. Our experience shows that it is numerically unstable in solving $\underline{\chi_j^2}$ whenever the computed $\underline{\chi_j^2}$ value is negative. Therefore it is better set any negative lower limit to zero, because χ_j^2 is always positive, unless all the principal components are zero, which is almost impossible. The lower limits ($\underline{\chi_j^2}$'s) should be obtained in reverse order, i.e., $\cdots, \underline{\chi_{j+1}^2}, \underline{\chi_j^2}, \underline{\chi_{j-1}^2}, \cdots$. Each can be used as an initial guess for the next to avoid numerical instability. $\underline{\chi_1^2}$ could be used as an initial guess for $\overline{\chi_1^2}$. The typical lower and upper limits for χ_j^2 are shown in Figure 6.9.

When the sequential chi-square test is used, for a given sample size of j, a sample value of $\chi_j^2 \le \underline{\chi_j^2}$ results in the acceptance of H_0, a sample value of $\chi_j^2 \ge \overline{\chi_j^2}$ results in the rejection of H_0, and a sample value of $\underline{\chi_j^2} < \chi_j^2 < \overline{\chi_j^2}$ results in an additional sample being taken.

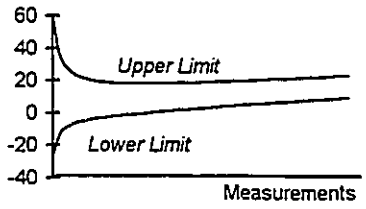


Figure 6.9 Upper and Lower Limits for
Sequential χ^2 Test

The collective sequential statistical test in general requires more samples in accepting the null hypothesis, because it takes some time before the lower limit exceeds zero.

The lower limit $\underline{\chi_j^2}$ and upper limit $\overline{\chi_j^2}$ for χ_j^2 can only be obtained if the generalized hypergeometric series in ${}_0F_1$ converges. The series does converge. In fact, there is a more general theorem which says that any generalized hypergeometric series, ${}_pF_q\left(\alpha_1,\cdots,\alpha_p;\beta_1,\cdots,\beta_q;z\right) = \sum_{k=0}^{\infty} \frac{(\alpha_1)_k \cdots (\alpha_p)_k}{(\beta_1)_k \cdots (\beta_q)_k} \frac{z^k}{k!}$, where $\alpha_1,\cdots,\alpha_p$; $\beta_1,\cdots,\beta_q$ are parameters, will converge for any finite z when $p \le q$ (Manual of Mathematics, 1979).

6.4.4 Choice of λ_1^2

There are a few ways to choose a value for λ_1^2 (Jackson and Bradley, 1961b). In general, correlation among variables affects the value of λ_1^2. Since principal components are not correlated, we may have an easier way to do this. Again we use the Sunoco Hydrocracker Fractionation Plant example to illustrate the methods. Assume the overall type I error is 0.05 and the tolerance for each principal component constitutes limits of $\pm 3\sigma\,(\sigma = 1)$ about their expectation (0). This corresponds to $\alpha = 0.0085$ for each principal component. As shown in Figure 6.10, the requirement that each principal component must be within its tolerance implies the true mean for a principal component cannot be closer than 2.631 to either tolerance limit; conversely, the true mean must be within 3 - 2.631 = 0.369 of the origin since the tolerances were assumed to be $\pm 3\sigma$ limits.

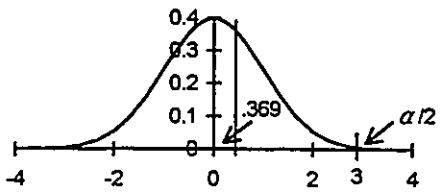


Figure 6.10 Normal distribution of a PC with mean 0

A schematic graph shown in Figure 6.11 illustrates two methods for defining λ_1^2. One method would involve inscribing a sphere inside the rectangular solid bounded by ± 0.369 for the 6 principal components. The radius under H_1 for a given principal component would be 0.369, therefore $\lambda_1^2 = 0.369^2 = 0.1362$.

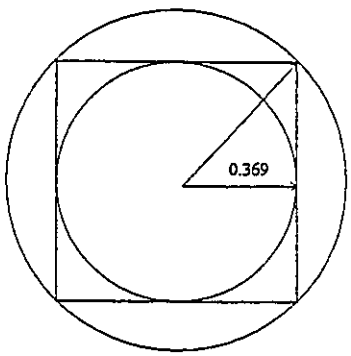


Figure 6.11 Methods to Determine λ_1^2

Another method would be to circumscribe a sphere around this rectangular solid. This would yield a value of $\lambda_1^2 = 6*0.369^2 = 0.8170$.

Considering the crudeness in determining λ_1^2 in the first place, there would be no definite preference for its value. This is rather a drawback of the methods.

6.4.5 Example: Sunoco Hydrocracker Fractionation Plant

Collective sequential analysis is most useful in on-line process monitoring and real-time data reconciliation. This is because it can provide us a single chart as opposed to a number of charts.

Figures 6.12 and 6.13 show the results of the sequential chi-square test for the principal components when data are arranged in the morning - evening order (A, B, C, D, E, F) and in the chronological order (A, D, B, E, C, F), respectively. Figure 6.12 shows one round of the sequential test using the data set A, B, C, and D[6], which detected the gross error. Figure 6.13 shows three rounds of the sequential test involving the pairs of the data set (A, D), (B, E), and (C, F). In the first two rounds, the test detected the gross error, while in the third round, more data were needed before a definite inference could be made.

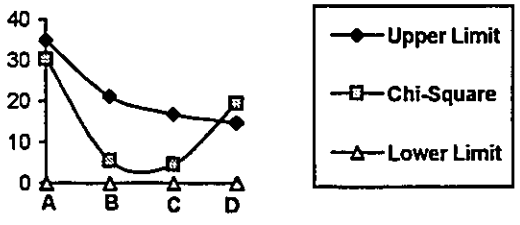


Figure 6.12 Sequential χ^2 Test for y_a, Data in Morning - Evening Order

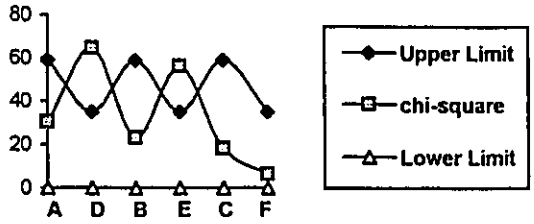


Figure 6.13 Sequential χ^2 Test for y_a, Data in Chronological Order

[6] E and F were omitted because they are not enough to finish the second round of the sequential test.

6.4.6 Example: Chemical Extraction Plant

Another example with many hours of data from a chemical extraction plant is given in Chapter 7 (Figure 7.5 and Figure 7.9), there data are analyzed by the sequential χ^2 test. The persistent gross errors are easily detected.

6.5. Identifying A Gross Error

After a gross error has been detected by either the univariate sequential test or the sequential chi-square test for the principal components, it can be identified by the method of contribution analysis discussed in Chapter 4. The identification can only be done for a single set of measurements. Usually this would be a representative set of measurements in a round of the sequential tests. A simple way to choose such a representative set is to choose the one that gives the largest principal component score for that particular principal component. Having said that, we will not pursue this topic further since the procedure will simply repeat the one already given in Chapter 4. In the Sunoco Hydrocracker Fractionation Plant example, the reconciliation for every set of measurements passed all the statistical tests when $\tilde{x}_2$ was deleted.

CHAPTER 7

DETECTING GROSS ERRORS AND ZERO ACCUMULATION

In previous chapters, we have studied how to detect and identify gross errors, and how to reconcile process measurements. However, we have not formally defined a criterion under which the so-called steady state data reconciliation method could be used.

A steady state process is a process whose state variables do not change values in time. A stationary process is a process in which the expectations of its state variables do not change values in time. A constant accumulation process is a process in which the expectations of the residuals of the process constraints do not change values in time. A special case of constant accumulation is zero accumulation. There is no mass accumulation in a zero accumulation process. When energy balances are considered, there is no energy accumulation in the process.

Another interesting problem, with less exposure in literature, is to detect and distinguish gross errors from periods of non-steady state. This problem must be solved because the data reconciliation method that we studied is valid only for a steady state process. In the conventional approach, departure from steady state is a type of gross error. However, steady state data reconciliation should not be done in the first place for a period of non-steady state[1]. The distinction is rather difficult because gross errors and non-steady states are all captured as outliers and thus confounded to a certain extent. It was common

[1] Except for a period of zero accumulation.

to assume that there was no gross error in measurements so that non-steady state could be easily detected. Nevertheless, the subjective assumption of no gross error is questionable and usually not justified.

In this chapter, we will establish the criterion under which the steady state data reconciliation method can be used, and present a method to determine when the criterion is met. This criterion is zero accumulation. A two-stage approach is proposed for distinguishing between zero accumulation and gross errors. A numerical example of a chemical extraction plant will be used to demonstrate the algorithms.

7.1. Steady State and Dynamic Data Reconciliation

The philosophy of data reconciliation is to adjust data to satisfy process constraints (conservation laws and other process equations). In a steady state process, the constraints are known *a priori* and assumed to be error free, while the measurements are stochastic and considered to be uncertain.

Unfortunately, things become much more complicated when a dynamic process is involved. Parameters and coefficients in the dynamic process constraints are also stochastic. Furthermore, the structure of a dynamic model may be uncertain in many cases as well. These uncertainties in model structure and parameters may be referred to as model uncertainty.

The dual uncertainties in data and model in a dynamic process make dynamic data reconciliation difficult. It is not possible in general to obtain a reliable reconciliation by adjusting data against a model in error, unless one could assume that the uncertainty in the

model could be safely ignored when compared to the uncertainty in the data, or one is dealing with a very simple process such as a stirred tank[2].

Some researchers have presented a few algorithms on dynamic data reconciliation (Almasy, 1990; Darouach and Zasadzinski, 1991). They all implicitly assumed that the dynamic models that they employed were error free. When a dynamic model is so assumed, dynamic data reconciliation can be reduced to a Kalman smoother or extended Kalman smoother problem[3]. However, the assumption itself may not hold in practical applications. A detailed discussion on dynamic data reconciliation is beyond the scope of this thesis. Nevertheless, we may understand from the earlier discussion that it is very important to define when the steady state data reconciliation method can be used.

7.2. Steady State, Stationary and Zero Accumulation Processes

In the data reconciliation literature, the terms *steady state process* and *stationary process* are used interchangeably. In fact, a steady state process is a special case of a stationary process. Both a steady state and a stationary process are a zero accumulation process. The converse may not be true. A zero accumulation process that is not a steady state or a stationary process is a dynamic process. For example, flow rates of a fast process may change values simultaneously without time delay. No accumulation occurs in the process even though the process may undergo dynamic changes.

It is clear that the criterion under which the steady state data reconciliation method can be used is zero accumulation, because the steady state model holds when there is no accumulation occurs in a process. The steady state data reconciliation method can be

[2] The model of a stirred tank is essentially theoretical.

[3] When the standard state space representation and the Kalman filtering cannot be applied, Darouach and Zasadzinski (1991) proposed a method using a singular model.

applied not only to a steady state process but also to a zero accumulation process. Therefore we may consider detecting zero accumulation as a generalization of detecting steady state, as far as data reconciliation is concerned.

The residuals of the reduced process constraints are the fundamental quantities in data reconciliation. These quantities can be utilized in detecting periods of zero accumulation and gross errors.

7.3. Problems in Steady State Detection

Narasimhan *et al.* (1986) proposed a test procedure for detecting periods of stationary state. They assumed that the process was in stationary state for a period of N sets of measurements, where N was fixed for all the periods. They tested whether measurements from two successive periods had the same means. The measurements were assumed to be statistically independent with no correlation with time. The test was applied to a subset of p variables chosen from all the variables, where the p variables were assumed to change state simultaneously in the same fashion. They further assumed that once the subset of p variables was chosen, any remaining variables were dismissed from further consideration. The problems with this procedure is that N cannot be known *a priori* and the same N cannot be appropriate in general to all the periods. The choice of p is rather subjective. It is unlikely in general that the same subset of the p variables can represent the changes in state of the process from time to time. Furthermore, the correlation of measurements with time is usually inevitable.

Holly *et al.* (1989) used the Student t-Test to compare means of each measured variable in successive periods to find periods of stationary state, which they called the *apparent* steady state. This is also a fixed sample approach, where the number of measurements in each period was determined by trial and error.

The problems in the existing methods of stationary (steady) state detection are: 1) the assumption that there is no gross error in the measurements; 2) the tests are for the measurements themselves; and 3) the assumption that the measurements are not correlated with time. In fact, one could never know in advance if there is any gross error in the measurements, nor the true means of the measurements. The measurements are always correlated with time. Furthermore, gross errors and non-stationary state are confounded and cannot be distinguished when the measurements themselves are directly tested.

Since the methods are based on a fixed number of samples in each period, they can not be used in on-line analysis without significant delay.

7.4. Detecting Gross Errors and Zero Accumulation

We will present a two-stage approach to detecting zero accumulation in the presence of gross errors based on sequential analysis. The assumption of no gross error is not needed. Gross errors will instead be detected if they exist. To accomplish this, the method does not test the measurements directly. It tests the residuals of the reduced process constraints e, whose expectations are known to be zero in the absence of gross error. The sample size of each period is not fixed. The method utilizes every measurement as soon as it becomes available and is capable of providing an early detection. The method can be used in on-line process monitoring and data reconciliation.

The first stage of the method detects if there is any period of zero accumulation. The second stage detects if there is any gross error in the measurements in a period of zero accumulation. Data reconciliation can be performed for all the periods of zero accumulation.

7.4.1 Sequential Chi-Square Test for Detecting Zero Accumulation

We apply the sequential chi-square test to detect zero accumulation. It was used by Jackson and Bradley (1961b) to test if the means of measured variables equal certain fixed values. The measurements were assumed not to be in gross error. We will compare the means of the residuals e in two successive periods where the measurements may be subject to gross error. The approach is closely related to that given in Chapter 6.

For each period determined by a sequential test, we test the null hypothesis H_0 that $E(e^i) = E(e^{i+1})$ against the alternative hypothesis H_1 that $E(e^i) \neq E(e^{i+1})$, $i = 1,2,\cdots$. If $E(e^i) = E(e^{i+1}) = c$, $c \neq 0$, the process may have constant non-zero accumulation or zero accumulation with data in gross error. This actually tests constancy of e and cannot distinguish between constant non-zero accumulation and zero accumulation unless $c = 0$. Therefore $E(e^i) = E(e^{i+1})$ is a necessary but not a sufficient test for zero accumulation.

In practice, every chemical unit has a fixed capacity. Constant non-zero accumulation cannot last forever or a process unit may overflow or run dry. It is then reasonable to assume that $c \neq 0$ is caused by gross errors, and constant accumulation suggests zero accumulation. In particular, if we start the test from a steady state window, the process naturally has zero accumulation.

When the sequential chi-square test is used, the null hypothesis is that $E(e^{i+1} - e^i)^T H_e^{-1} E(e^{i+1} - e^i) = \lambda_0^2 = 0$ and the alternative hypothesis is that $E(e^{i+1} - e^i)^T H_e^{-1} E(e^{i+1} - e^i) = \lambda_1^2 \neq 0$.

The algorithm for detecting constant accumulation is as follows:

A) Prepare tables for the lower limit $\underline{\chi_j^2}$ and upper limit $\overline{\chi_j^2}$ of the chi-square variable χ_j^2 by solving the set of the non-linear equations

$$ {}_0F_1\left(\frac{m}{2}, \frac{j\lambda_1^2\underline{\chi_j^2}}{4}\right) exp\left(\frac{-j\lambda_1^2}{2}\right) = \frac{\beta}{1-\alpha} \tag{7-1} $$

$$ {}_0F_1\left(\frac{m}{2}, \frac{j\lambda_1^2\overline{\chi_j^2}}{4}\right) exp\left(\frac{-j\lambda_1^2}{2}\right) = \frac{1-\beta}{\alpha} \tag{7-2} $$

where j is the number of the sets of measurements used in the current sequential test, $j = 1,2,\cdots$

B) Select α, β, and λ_1 as outlined in Chapter 6

C) Choose a base period (e.g., a steady state window) that represents the nominal operation of the process from which the test will start, and obtain the mean of the base period

D) Set the counter variable $j = 1$

E) For the j-th set of the samples $\tilde{x}^j$, compute the residual vector e^j and the chi-square variable $\chi_j^2 = j\left(\bar{e} - \bar{e}_p\right)^T H_e^{-1}\left(\bar{e} - \bar{e}_p\right)$, where $\bar{e}$ is the vector of sample means based on all of j observations, $\bar{e} = \frac{1}{j}\sum_{i=1}^{j} e^i$; $\bar{e}_p$ is the vector of sample means of the earlier period. If this is the very first round of the test, $\bar{e}_p$ is obtained from the base period ($\bar{e}_p$ may not be zero if gross errors are present)

F) If $\chi_j^2 \le \underline{\chi_j^2}$, accept H_0. Assign $\bar{e}_p = \bar{e}$. Flag the current and the earlier periods as constant accumulation. Go to D) until no more data left

If $\chi_j^2 \ge \overline{\chi_j^2}$, reject H_0. Assign $\bar{e}_p = \bar{e}$. If the earlier period was flagged as constant accumulation, flag that period as the end of constant accumulation; otherwise flag that period as non-constant accumulation. Go to D) until no more data left

If $\underline{\chi_j^2} < \chi_j^2 < \overline{\chi_j^2}$, more samples are needed. Increase j by 1 and go to E) until no more data left

7.4.2 Sequential Chi-Square Test for Detecting Gross Errors

We are already familiar with the fact that the expectations of the residuals of the reduced process constraints, $E(e)$, are zero when there is no gross error. For the i-th

period determined by a sequential test, we test the null hypothesis that $E(e^i) = 0$ against the alternative hypothesis that $E(e^i) \neq 0$, $i = 1,2,\cdots$. In particular, when the sequential chi-square test is used, the null hypothesis is that $E(e^i)^T H_e^{-1} E(e^i) = \lambda_0^2 = 0$ and the alternative hypothesis is that $E(e^i)^T H_e^{-1} E(e^i) = \lambda_1^2 \neq 0$.

The algorithm for detecting gross errors is as follows:

A) Prepare tables for the lower limit $\underline{\chi_j^2}$ and upper limit $\overline{\chi_j^2}$ of the chi-square variable χ_j^2 by solving the set of the non-linear equations (7-1) and (7-2), where j is the number of the sets of measurements used in the current sequential test, $j = 1,2,\cdots$

B) Select α, β, and λ_1 as outlined in Chapter 6

C) Set the counter variable $j = 1$

D) For the j-th set of the samples $\tilde{x}^j$, compute the residual vector e^j and the chi-square variable $\chi_j^2 = j\bar{e}^T H_e^{-1} \bar{e}$, where $\bar{e}$ is the vector of sample means based on all of j observations, $\bar{e} = \frac{1}{j}\sum_{i=1}^{j} e^i$

E) If $\chi_j^2 \leq \underline{\chi_j^2}$, accept H_0. Flag the period as having no gross error. Go to C) until no more data left

If $\chi_j^2 \geq \overline{\chi_j^2}$, reject H_0. Flag the period as having gross error. Go to C) until no more data left

If $\underline{\chi_j^2} < \chi_j^2 < \overline{\chi_j^2}$, more samples are needed. Increase j by 1 and go to D) until no more data left

7.5. Example: Chemical Extraction Plant

7.5.1 Case 1

Holly *et al.* (1989) presented an example of a chemical extraction plant, shown in Figure 7.1. Data of total flow rates were available over a 48 hour period, with

measurements in 22 measured streams every 6 minutes. Values of the total flow rates against time are shown in Figure 7.2, along with some comments on the measurements in Table 7.1. One stream (F106) had constant zero flow rate and is not shown in Figure 7.2.

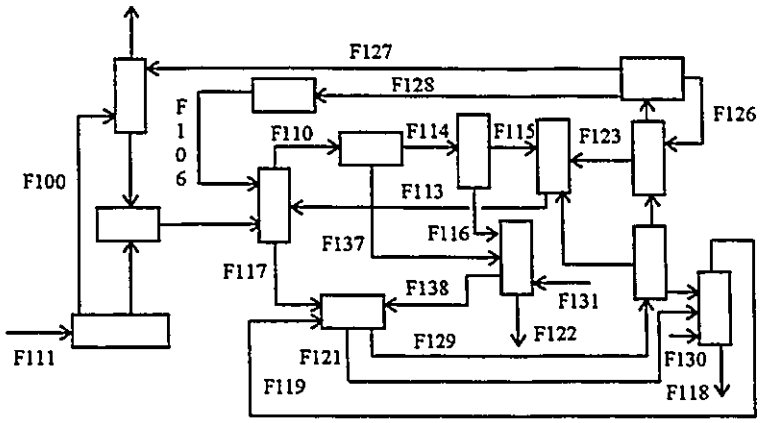
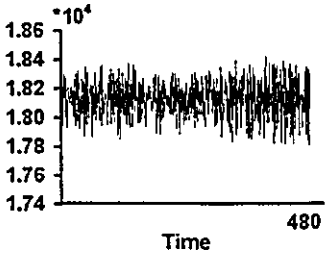
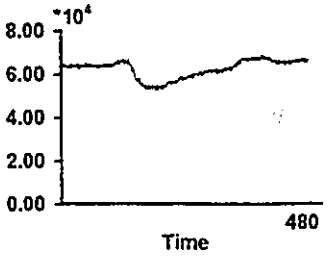


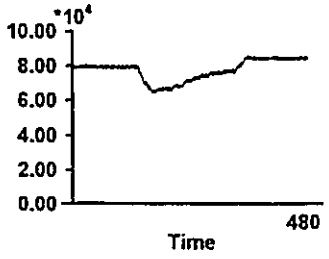
Figure 7.1 Chemical Extraction Plant



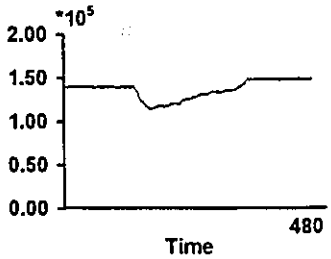
F100



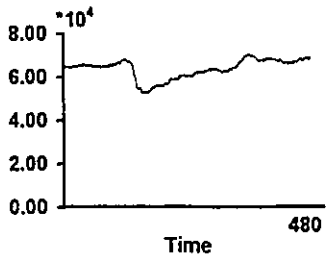
F110



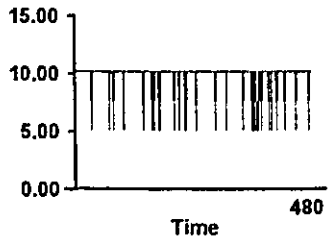
F111



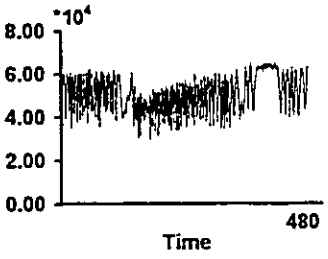
F113



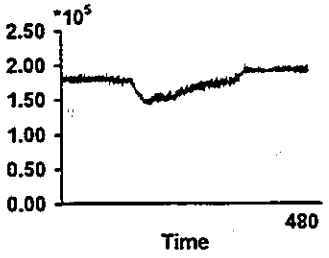
F114



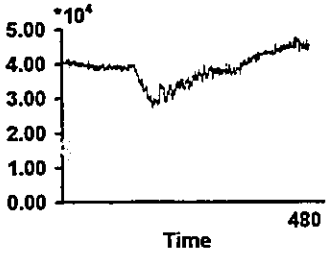
F115



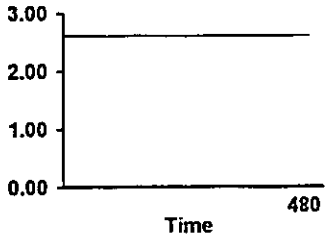
F116



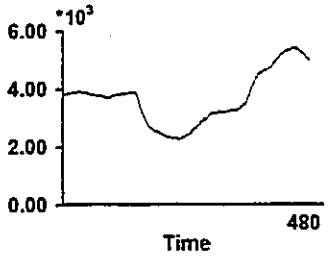
F117



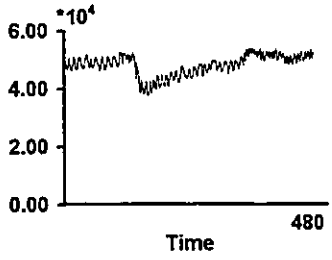
F118



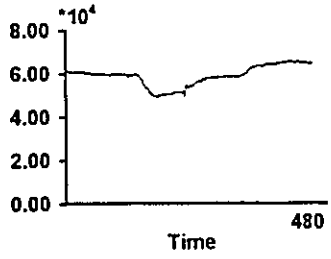
F119



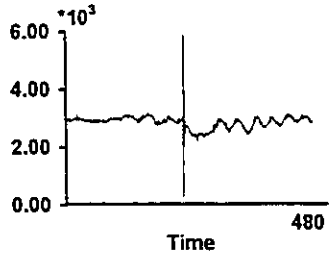
F121



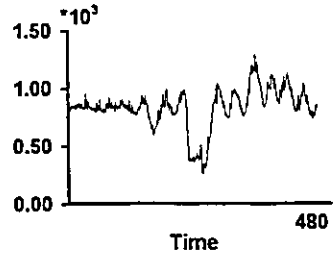
F122



F123



F126



F127

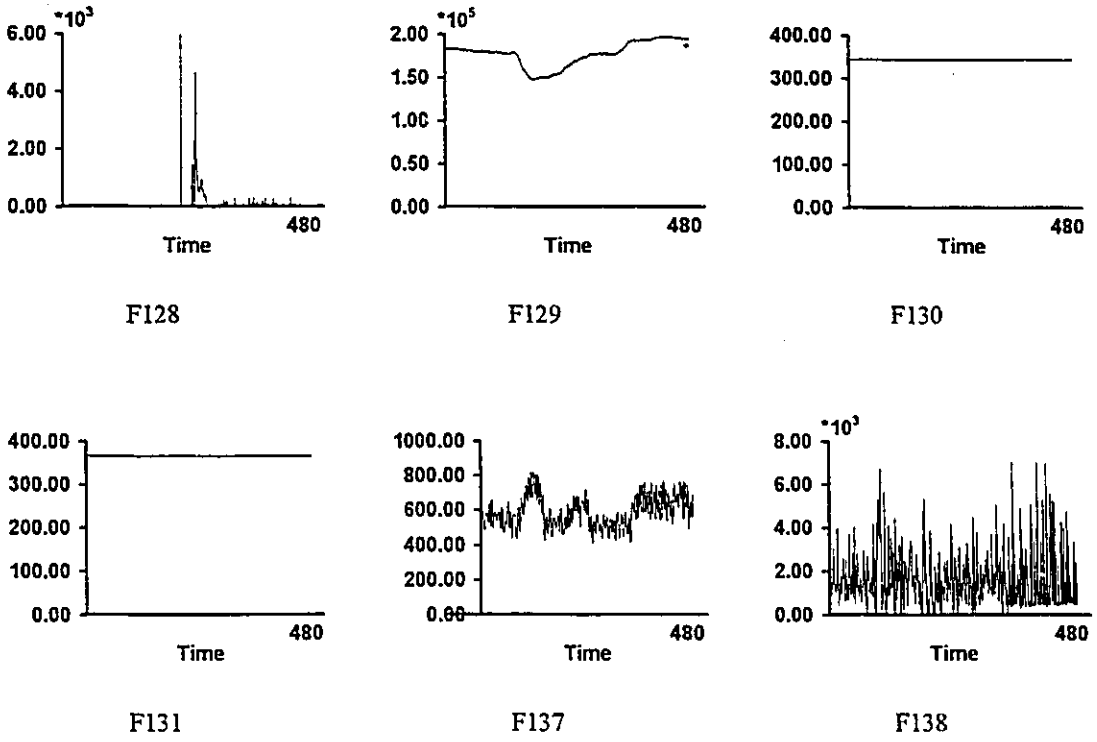


Figure 7.2 Flow Rate Measurements, Case 1

By inspecting the total flow rates shown in Figure 7.2, we noticed that there were different patterns of the flow rates: stationary-dynamic-stationary, stationary-dynamic, stationary with a few spikes, stationary, non-zero constant with some spikes, non-zero constant, and zero constant. Holly *et al.* concluded based on the Student t-Test that the first 130 sets of measurements (the first 13 hours) were taken from an *apparent* steady state.

There were 16 total mass balances. The reduced balances are shown in Table 7.2.

Table 7.1 Patterns of Flow Rates, Case 1

Stream Tag	Comments
F110	Stationary-dynamic-stationary
F111	Stationary-dynamic-stationary
F113	Stationary-dynamic-stationary
F114	Stationary-dynamic-stationary
F116	Stationary-dynamic-stationary
F117	Stationary-dynamic-stationary
F122	Stationary-dynamic-stationary
F123	Stationary-dynamic-stationary
F127	Stationary-dynamic-stationary
F129	Stationary-dynamic-stationary
F137	Stationary-dynamic-stationary
F118	Stationary-dynamic
F121	Stationary-dynamic
F126	Stationary with a few spikes
F100	Stationary
F138	Stationary
F115	Constant with some spikes
F128	Constant with some spikes
F119	Constant non-zero
F130	Constant non-zero
F131	Constant non-zero
F106	Constant zero (removed)

Table 7.2 Reduced Balances

1	F111 - F118 - F122 = 0
2	F114 - F115 - F116 = 0
3	F116 - F122 + F131 + F137 - F138 = 0
4	F110 - F114 - F137 = 0
5	F117 + F119 - F121 - F129 + F138 = 0
6	-F113 + F115 - F118 - F119 + F121 - F127 - F128 + F129 + F130 = 0

The normalized residuals of the reduced constraints are plotted in Figure 7.3. Their thresholds are ±2.631 for 95% overall confidence level and 5 degrees of freedom. It is interesting to observe that 1) all residuals are biased from zero with those in e_3 less

noticeable; 2) all residuals are well within their upper and lower thresholds, except e_2; 3) the outliers in e_2 are well scattered. This suggests that if each set of measurements were tested alone, we would not have been able to detect the dynamic changes in the process. Most of time we would not be able to detect any gross error[4].

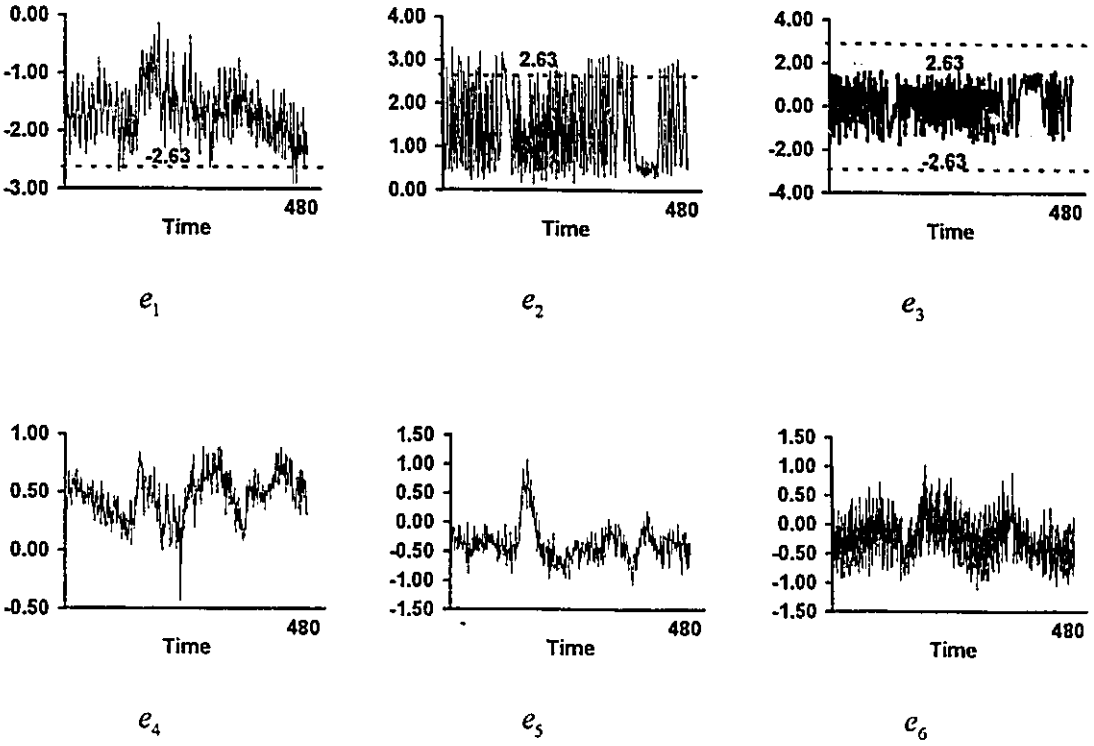


Figure 7.3 Normalized Residuals of the Reduced Constraints, Case 1

[4] There are 58 outliers out of the 480 residuals (12.1%) for the second reduced constraint (e_2). This means that there would be 87.9% of time that we could not detect the gross error if a non-sequential test for e_2 were used.

7.5.1.1 Detecting Constant Accumulation

The result of detecting constant accumulation is given in Figure 7.4, with $\alpha = 0.05$, $\beta = 0.001$, and $\lambda_1^2 = 0.82$. Figure 7.4 shows the lower limit ($\underline{\chi_j^2}$'s) and upper limit ($\overline{\chi_j^2}$'s) of the sequential chi-square statistics (χ_j^2's). Initially, the upper limit is at its highest value and the lower limit is at its lowest value. As more data become available, the upper limit quickly moves down, and the lower limit moves up. The distance between the two limits becomes smaller (but one never intersects the other). Whenever a sequential chi-square statistic goes beyond either limit, the current sequential test ends and the next one starts. The upper limit is reset to its highest value, and the lower limit to its lowest value. 480 sets of data were examined. The null hypothesis that the process was at constant accumulation was accepted in the time intervals I_1:[1, 142] (about 14 hours), I_2:[176, 254] (about 8 hours), and I_3:[450, 480] (about 3 hours), and rejected elsewhere. We suggest that the periods I_1 and I_2 may be considered as zero accumulation since they are long enough. The first period of zero accumulation I_1 approximately matched the period of apparent steady state found by Holly *et al.* During the second interval I_2 the process was undergoing dynamic changes in such a manner that no mass accumulation occurred anywhere in the process. In the period I_3 some flow rates returned to stationary state while a few others were still dynamic. As a matter of fact, the sequential statistics in that interval were all within the upper and lower limits. It would make no difference either to accept or to reject the null hypothesis from the statistical point of view. However, considering that the

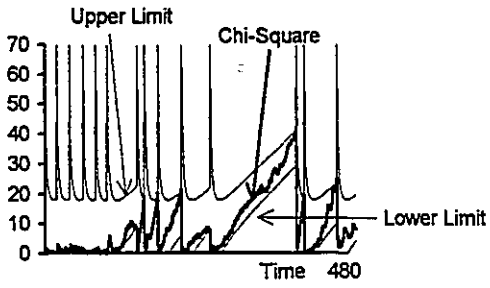


Figure 7.4 Sequential Test for Zero Accumulation

statistics were closer to the lower limit, we have flagged I_3 as a period of zero accumulation.

7.5.1.2 Detecting Gross Errors

Having flagged the periods of zero accumulation, we can proceed to detect if there is any gross error in these periods. The results are given in Figures 7.5.1 and 7.5.2[5], with the same values of α, β and λ_1^2. Again in these figures we show the lower limit ($\underline{\chi_j^2}$'s) and upper limit ($\overline{\chi_j^2}$'s) of the sequential chi-square statistics (χ_j^2's). Whenever a sequential chi-square statistic goes beyond either limit, the current sequential test ends and the next one starts. The upper limit is reset to its highest value, and the lower limit to its lowest value. The null hypothesis that there is no gross error was rejected by each and every sequential test, with only 2~3 sets of measurements in each test. This is very strong evidence of gross errors.

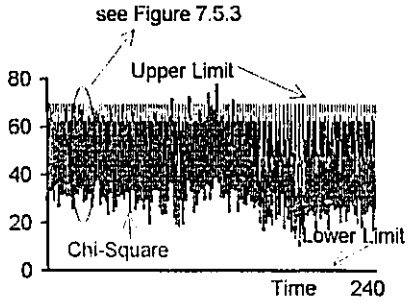


Figure 7.5.1 Sequential Test for Gross Errors

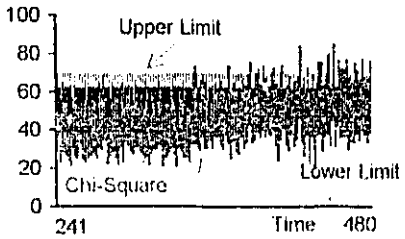


Figure 7.5.2 Sequential Test for Gross Errors

[5] We actually showed the statistics for all measurements, including those for the periods of non-zero accumulation. Therefore, the outliers in such periods may be also caused by the departure from steady state.

In the earlier section, it took more measurements in each test to detect zero accumulation. That was because the difference in the means of the successive periods was not as large as that of a period from zero.

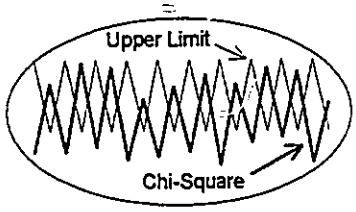
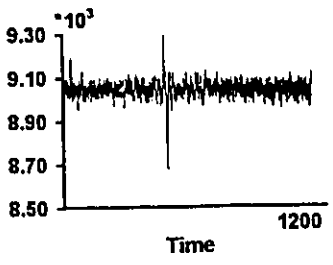


Figure 7.5.3 Enlarged Area of Figure 7.5.1

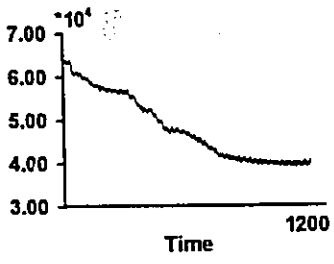
The method that we presented here can be used to quickly detect gross errors in the measurements. To identify where the gross errors are, we can use the contribution analysis discussed in earlier chapters. The topic has been exploited in great detail. We will not repeat it in this example.

7.5.2 Case 2

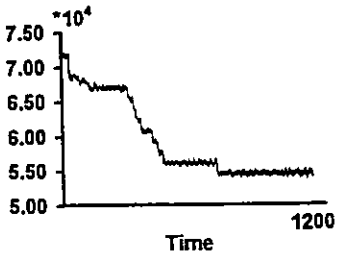
We obtained new data a few years later, from the same chemical extraction plant. In this new case, data of total flow rates were available over a 120 hour period. The covariance matrix changed reflecting the fact that the plant operation was at different level. The other conditions remained unchanged. Values of the total flow rates against time are shown in Figure 7.6, along with some comments on the measurements in Table 7.3.



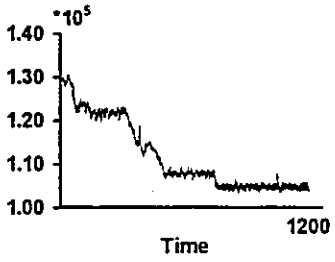
F100



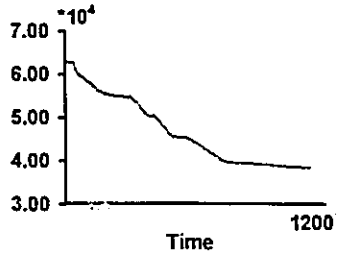
F110



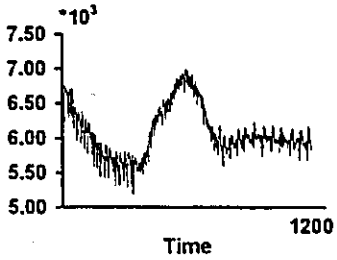
F111



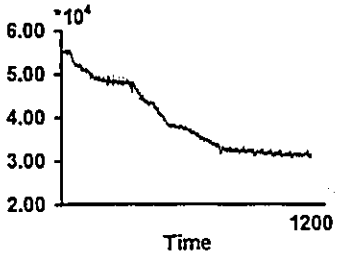
F113



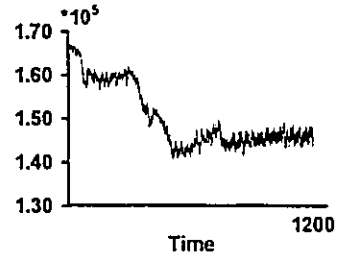
F114



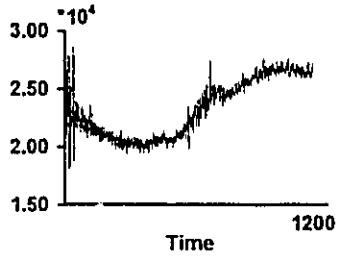
F115



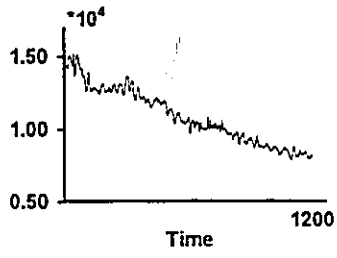
F116



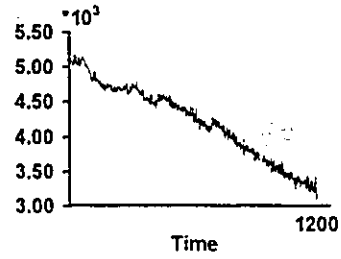
F117



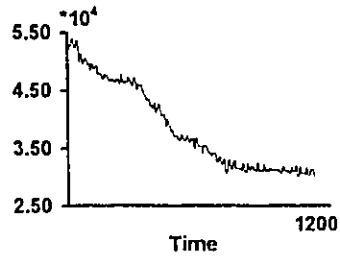
F118



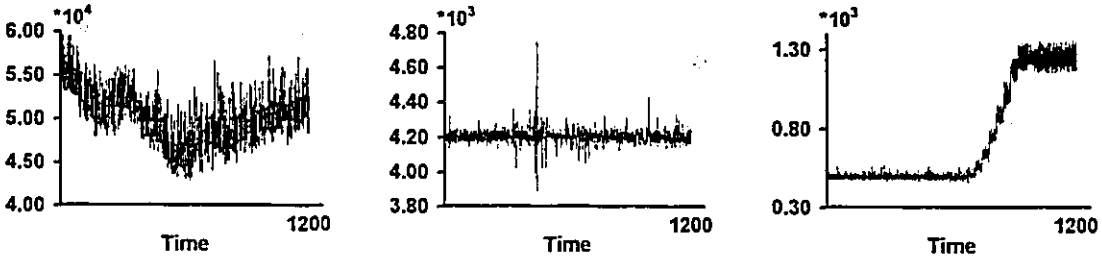
F119



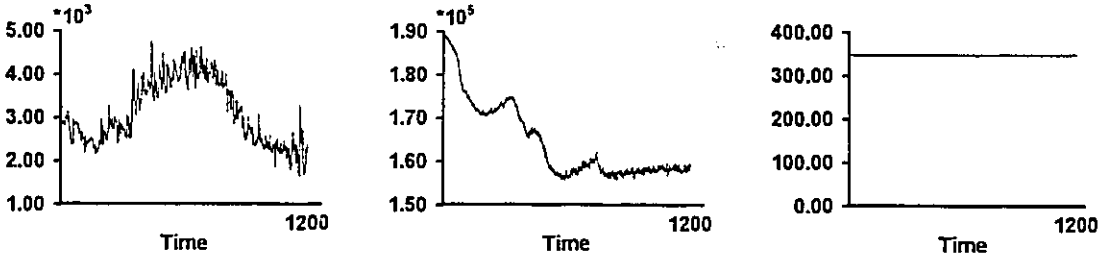
F121



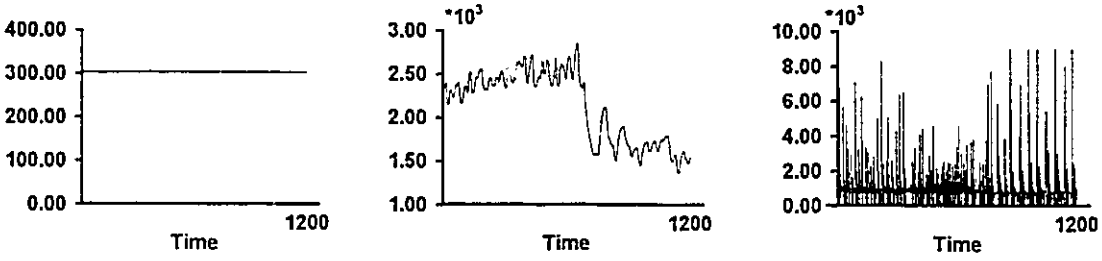
F122



F123 F126 F127



F128 F129 F130



F131 F137 F138

Figure 7.6 Flow Rate Measurements, Case 2

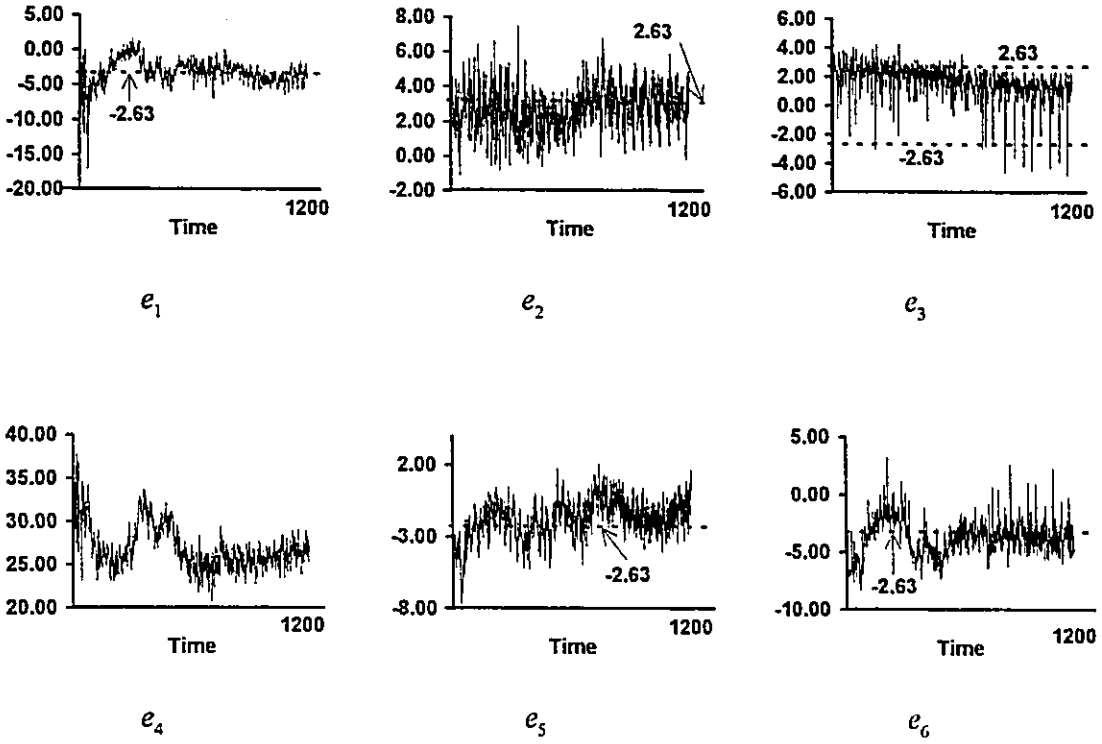


Figure 7.7 Normalized Residuals of the Reduced Constraints, Case 2

The major difference of this case from the other one is that the process was essentially dynamic. It is almost impossible to find out any period of zero accumulation by simply inspecting the flow rates by eye. Figure 7.7 shows the normalized residuals of the reduced constraints. Many of them are outliers.

The result of detecting zero accumulation is shown in Figure 7.8, with $\alpha = 0.05$, $\beta = 0.001$, and $\lambda_1^2 = 0.82$. The intervals I_1:[791, 857] (6.6 hours) and I_2:[981, 1065] (8.4 hours) were flagged as the periods of zero accumulation. We should note that in those two periods the process was still dynamic. However, data reconciliation can be done by using the steady state data reconciliation method.

Figure 7.9 shows the result of detecting gross errors, with the same values of α, β and λ_1^2. All the statistics were much larger than the upper limit, showing very strong evidence of gross error and dynamics as well.

Table 7.3 Patterns of Flow Rates, Case 2

Stream Tag	Comments
F100	Stationary with a few spikes
F126	Stationary with a few spikes
F138	Stationary with spikes
F127	Stationary-dynamic-stationary
F110	Dynamic
F114	Dynamic
F116	Dynamic
F118	Dynamic
F119	Dynamic
F121	Dynamic
F123	Dynamic
F128	Dynamic
F111	Dynamic-stationary
F113	Dynamic-stationary
F115	Dynamic-stationary
F117	Dynamic-stationary
F122	Dynamic-stationary
F129	Dynamic-stationary
F137	Dynamic-stationary
F130	Constant non-zero
F131	Constant non-zero
F106	Constant zero (removed)

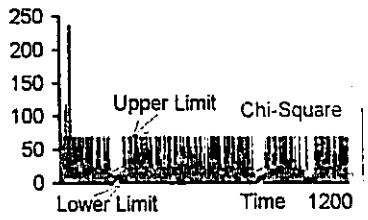


Figure 7.8 Testing for Zero Accumulation

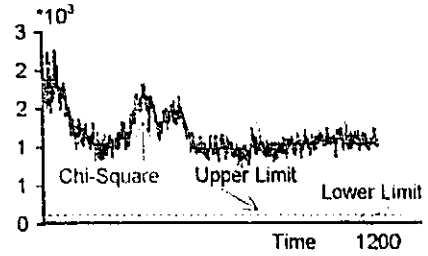


Figure 7.9 Testing for Gross Errors

7.6. Limitation of the Method and Future Work

Sequential chi-square test is effective and efficient in detecting gross errors and zero accumulation. The computations involved are actually simple matrix multiplication after the tables of lower and upper limits have been prepared. Since the tables can be prepared off-line, the method can be used on-line without heavy computations.

Though we used $\boldsymbol{e}$ in the sequential chi-square test, we actually implicitly used all principal components y_e in the test, because a chi-square variable can be equivalently expressed in either $\boldsymbol{e}$ or y_e, as we have seen in Chapters 3 and 4. Therefore it is natural to consider if a sequential truncated chi-square test can be implemented.

The algorithm can be easily modified to perform sequential truncated chi-square test for testing one of the null hypotheses of $E(y_e) = 0$, $E(y_r) = 0$, or $E(y_a) = 0$, with the numbers of retained principal components, k_e, k_r and k_a respectively. In each case, H_e^{-1} is replaced by an identity matrix of appropriate size, since the covariances of

principal components are all identity matrices. It should be noted that there is no direct sequential chi-square test for a and r since H_r and Q_1 are singular.

However, one must be cautious if one wants to use a sequential *truncated* chi-square test. Sometimes the unretained principal components might contain important information regarding gross errors and non-zero accumulations. This would not be a problem in non-sequential truncated chi-square tests, because the Q statistic is available to take care of the unretained principal components. Unfortunately, the sequential Q test has not been established so far. We hope that future research will lead to a development of such a test, so that on-line computations can be done in reduced dimensions.

CHAPTER 8

SUMMARY AND CONCLUSIONS

8.1. Thesis Contributions

We have studied some important issues in data reconciliation that process contaminated data into consistent information. The major contributions of this work are summarized below.

A modification of the model of bilinear data reconciliation using Lagrange multipliers was developed, which is suited to reconcile measurements with an arbitrary covariance structure. In addition, it was proved that the ranks of the variance-covariance matrices used in data reconciliation are all identical and could be computed in advance.

The quality of data reconciliation depends upon the accuracy of the model. A model that consists of steady state material and energy balances is best suited to reconcile primary variables such as total flow rates and component flow rates. The steady state data reconciliation method can be used to reconcile data from a steady state process or a process with zero accumulation. Dynamic data reconciliation can only be done when the structure and the parameters of the dynamic model are accurate. This is usually a Kalman smoother problem.

Gross errors must be detected and identified. The principal component tests based on PCA were developed. The tests are sharper in detecting, and have a substantially greater power in correctly identifying, the gross errors than do the currently used statistical tests in data reconciliation. The tests are helpful in preventing us from wrongly deleting good measurements while keeping corrupted ones.

The power of the PC tests can be explained by two reasons. The first reason is due to the principal component transformation, which transfers correlated original variables into uncorrelated PCs, so that the derived tests are sharper and less confounding. The second reason is due to the combination of PCA and the steady state model. Unlike conventional PCA applications in chemical engineering, which directly use original data, the PC tests start from $\boldsymbol{e}$, $\boldsymbol{r}$ and $\boldsymbol{q}$, and take advantage of the prior knowledge about them. This is because it is impossible to know the true expectations of the original variables, while the true expectations of $\boldsymbol{e}$, $\boldsymbol{r}$ and $\boldsymbol{q}$ are zeros.

A difficulty in performing PCA on $\boldsymbol{r}$ and $\boldsymbol{q}$ is the singularity of the variance-covariance matrices. The problem was solved by the theoretical proof of the ranks of those matrices. This gives the maximum number of PCs. The relationship between the PC test and the other conventional tests was discussed. It was shown that the univariate, MP and PC test are identical under certain limiting condition, but in general the univariate and the MP tests cannot compete with the PC tests.

It was demonstrated that a MP test could be misleading. In a bilinear problem, the MP measurement test should be performed on $\boldsymbol{q}$, not on $\boldsymbol{a}$ and $\boldsymbol{\delta}$ separately, when $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{d}}$ are correlated.

To detect persistent gross errors, Sequential Analysis was applied to the principal components to test them both individually and collectively. The number of measurements required in a sequential test is not determined in advance but is dependent on the outcome of the measurements as they are made. A key merit of this procedure is that it requires in general many fewer measurements than other procedures based on a fixed number of measurements. It is shown that the collective sequential test can only be applied to the principal components, not to the original optimal measurement adjustments, because of the singularity in the covariance matrix of the adjustments.

A two-stage test was presented to distinguish between zero accumulation and gross error. This allows us to determine the applicability of the steady state data reconciliation model and algorithms, and perform sequential tests with measurements in gross error.

The use of PCA makes it possible to calculate the exact probability of type I error for a PC and of the overall type I error for the PCs.

8.2. Future Work

We observed an interesting phenomenon from the numerical examples that H_e in general is not well conditioned, and tends to be extremely ill-conditioned as a problem becomes larger and complicated. Though H_e is of full rank, it may become numerically singular during iterations. This may cause a problem not to converge in bilinear data reconciliation. The exact reason of the ill-conditioning is unknown at this time. A better understanding of the nature of H_e would lead to a more robust bilinear data reconciliation algorithm.

In contribution analysis, the tolerances ε_1 and ε_2 are used in determining the number of major contributors to an inflated PC statistic. Further theoretical and industrial studies are needed to establish a suitable range for such tolerances.

Perhaps our work in detecting zero accumulation is only a start. We anticipate that further studies may result in more elaborated algorithms to distinguish between gross errors and zero accumulation.

In this thesis, we touched the issue of dynamic data reconciliation. Although it is not our theme, our study indicated that the success of data reconciliation depends upon its model. In particular, the success of dynamic data reconciliation depends upon a good dynamic model.

As we pointed out, dynamic data reconciliation can be reduced to a Kalman smoother problem when the model is correct. A promising direction is to utilize the information available from the Kalman smoother and PCA to derive a set of statistical tests to detect gross errors in dynamic data reconciliation. The tests should be analogous to what we have developed in this thesis, a combination of PCA and Kalman smoother. We are working towards this direction. The preliminary results are encouraging.

The implementation of data reconciliation in industries has been behind that of simulation and optimization, despite the fact that good data are crucial for simulation and optimization. This situation was caused by the lack of accuracy in estimating the measurement variance-covariance matrix and the difficulty in identifying gross errors. The use of PCA and contribution analysis is helpful in coping with the gross error identification problem. We hope that this method can be further implemented in large-scale industrial problems.

APPENDIX A

RELATIONSHIPS IN LINEAR DATA RECONCILIATION

Table A.1 lists important variables in linear data reconciliation. Vector $\tilde{x}$ consists of the measurements which provide original information regarding the process under investigation. Vectors e and r consist of the residuals of the process constraints. Together with a, they show how consistent the measurements are. Vectors λ and μ are the Lagrange multipliers for the constraints. They are not only directly related to the MP statistics, but also used in compactly expressing e, r and a.

Table A.1 shows that all the variables are related. It is interesting to see how the measurements ($\tilde{x}$), instrument precision (Σ_1), process flowsheet structure (B_1), and sensor placement (Y) play their roles in data reconciliation. The relationships are useful in theoretical analysis such as in the computation of the ranks of the variance-covariance matrices in hypothesis testing. In addition, there are uniform expressions for e, r and a in both linear and bilinear data reconciliation. These uniform expressions, shaded in Table A.1, can make a computer code more compact and efficient.

Table A.2 lists the variance-covariance matrices used in hypothesis testing. Table A.3 summarizes the assumptions used in the statistical tests. All of the assumptions were derived from that the measurements follow a multivariate normal distribution with a known variance-covariance matrix.

Table A.2 and Table A.3 would be identical respectively to Table B.2 and Table B.3, given in Appendix B, if the matrices used only in bilinear data reconciliation were eliminated from Table B.2 and Table B.3.

Table A.1 should be read from top to bottom, while Table A.2 and Table A.3 should be read from left to right.

Table A.1 Relationships Among the Variables[1]

	$e=$	$r=$	$a=$	$\lambda=$	$\mu=$
e	—	$H_0 Y H_e^{-1} e$	$-\Sigma_1 B_1^T Y H_e^{-1} e$	$-H_e^{-1} e$	$-Y H_e^{-1} e$
r	$Y^T r$	—	$-\Sigma_1 B_1^T Y H_e^{-1} Y^T r$	$-H_e^{-1} Y^T r$	$-Y H_e^{-1} Y^T r$
a	$-Y^T B_1 a$	$-B_1 a$	—	$H_e^{-1} Y^T B_1 a$	$Y H_e^{-1} Y^T B_1 a$
λ	$-H_e \lambda$	$-H_0 Y \lambda$	$\Sigma_1 B_1^T Y \lambda$	—	$Y\lambda$
μ	$-Y^T H_0 \mu$	$-H_0 \mu$	$\Sigma_1 B_1^T \mu$	$H_e^{-1} Y^T H_0 \mu$	—
$\tilde{x}$	$Y^T E_1$	$H_0 Y H_e^{-1} Y^T E_1$	$-\Sigma_1 B_1^T Y H_e^{-1} Y^T E_1$	$-H_e^{-1} Y^T E_1$	$-Y H_e^{-1} Y^T E_1$

Table A.2 Covariance Matrices

Covariance	Expression	Other Expression(s)	When $Y^T = I$
$H_e = cov(e)$	$Y^T H_0 Y$	$Y^T B_1 Q_1 B_1^T Y$; $Y^T H_r Y$	H_0; $B_1 \Sigma_1 B_1^T$; $B_1 Q_1 B_1^T$
$H_r = cov(r)$	$H_0 Y H_e^{-1} Y^T H_0$	$B_1 Q_1 B_1^T$	H_0; $B_1 Q_1 B_1^T$
$Q_1 = cov(a)$	$\Sigma_1 B_1^T Y H_e^{-1} Y^T B_1 \Sigma_1$	—	$\Sigma_1 B_1^T H_e^{-1} B_1 \Sigma_1$
$cov(\lambda)$	H_e^{-1}	—	H_e^{-1}
$cov(\mu)$	$Y H_e^{-1} Y^T$	—	H_e^{-1}
$cov(\Sigma_1^{-1} a)$	$\Sigma_1^{-1} Q_1 \Sigma_1^{-1}$	$B_1^T Y H_e^{-1} Y^T B_1$	$\Sigma_1^{-1} Q_1 \Sigma_1^{-1}$; $B_1^T H_e^{-1} B_1$

[1] We denote that $H_0 = B_1 \Sigma_1 B_1^T$, $H_e = Y^T H_0 Y$, and $E_1 = B_0 c + B_1 \tilde{x}$

Table A.3 Statistical Tests

Statistical Test	Based on
z_e	$e \sim N(0, H_e)$
z_r	$r \sim N(0, H_r)$
z_a	$a \sim N(0, Q_1)$
z_e^*	$\lambda \sim N\left(0, H_e^{-1}\right)$
z_r^*	$\mu \sim N\left(0, YH_e^{-1}Y^T\right)$
z_a^*	$\Sigma_1^{-1}a \sim N\left(0, \Sigma_1^{-1}Q_1\Sigma_1^{-1}\right)$ or $\Sigma_1^{-1}a \sim N\left(0, B_1^T YH_e^{-1}Y^T B_1\right)$
y_e, Q_e, and $\chi^2_{k_e}$	$e \sim N(0, H_e)$ and eigensystem of H_e
y_r, Q_r, and $\chi^2_{k_r}$	$r \sim N(0, H_r)$ and eigensystem of H_r
y_a, Q_a, and $\chi^2_{k_a}$	$a \sim N(0, Q_1)$ and eigensystem of Q_1

APPENDIX B

RELATIONSHIPS IN BILINEAR DATA RECONCILIATION

Appendix B parallels to Appendix A. Table B.1 shows the relationships among the variables in bilinear data reconciliation. Table B.2 lists the variance-covariance matrices used in hypothesis testing. Table B.3 summarizes the assumptions used in the statistical tests. All of the assumptions were derived from that the measurements follow a multivariate normal distribution with a known variance-covariance matrix. The expressions in those tables would reduce to their counterparts in Appendix A if the matrices used only in bilinear data reconciliation, such as Z, N, B_2 and $\tilde{d}$, are deleted. The results agree with those of Crowe (1986) when category 1 and category 2 variables are not correlated.

The uniform expressions of the variables in both linear and bilinear data reconciliation are shaded in Table B.1. Table B.1 should be read from top to bottom, while Table B.2 and Table B.3 should be read from left to right.

As a shortcut, the relationships among the variables in bilinear data reconciliation can be obtained from their linear counterparts shown in Appendix A by replacing Y with YZ, B_1 with $\begin{bmatrix} B_1 & B_2 N \end{bmatrix}$, Σ_1 with Σ, $\tilde{x}$ with $\begin{bmatrix} \tilde{x} \\ \tilde{d} \end{bmatrix}$, and a with $\begin{bmatrix} a \\ \delta \end{bmatrix}$.

Table B.1 Relationships among the Variables[1]

	$e=$	$r=$	$\begin{bmatrix} a^T & \delta^T \end{bmatrix}^T=$	$\lambda=$	$\mu=$
e	—	$H_0YZH_e^{-1}e$	$-\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZH_e^{-1}e$	$-H_e^{-1}e$	$-YZH_e^{-1}e$
r	Z^TY^Tr	—	$-\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZH_e^{-1}Z^TY^Tr$	$-H_e^{-1}Z^TY^Tr$	$-YZH_e^{-1}Z^TY^Tr$
$a,\ \delta$	$-Z^TY^TE_2$	$-E_2$	—	$H_e^{-1}Z^TY^TE_2$	$YZH_e^{-1}Z^TY^TE_2$
λ	$-H_e\lambda$	$-H_0YZ\lambda$	$\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZ\lambda$	—	$YZ\lambda$
μ	$-Z^TY^TH_0\mu$	$-H_0\mu$	$\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}\mu$	$H_e^{-1}Z^TY^TH_0\mu$	—
$\tilde{x},\tilde{d}$	$Z^TY^TE_1$	$H_0YZH_e^{-1}Z^TY^TE_1$	$-\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZH_e^{-1}Z^TY^TE_1$	$-H_e^{-1}Z^TY^TE_1$	$-YZH_e^{-1}Z^TY^TE_1$

Table B.2 The Covariance Matrices

Covariance	Expression	Other Expression(s)
$H_e=cov(e)$	$Z^TY^TH_0YZ$	$Z^TY^T\begin{bmatrix} B_1 & B_2N \end{bmatrix}Q\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZ$; $Z^TY^TH_rYZ$
$H_r=cov(r)$	$H_0YZH_e^{-1}Z^TY^TH_0$	$\begin{bmatrix} B_1 & B_2N \end{bmatrix}Q\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}$
$Q=cov\left(\begin{bmatrix} a \\ \delta \end{bmatrix}\right)$	$\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZH_e^{-1}Z^TY^T\begin{bmatrix} B_1 & B_2N \end{bmatrix}\Sigma$	—
$cov(\lambda)$	H_e^{-1}	—
$cov(\mu)$	$YZH_e^{-1}Z^TY^T$	—
$cov\left(\Sigma^{-1}\begin{bmatrix} a \\ \delta \end{bmatrix}\right)$	$\Sigma^{-1}Q\Sigma^{-1}$	$\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}YZH_e^{-1}Z^TY^T\begin{bmatrix} B_1 & B_2N \end{bmatrix}$

[1] We denote that $H_0=\begin{bmatrix} B_1 & B_2N \end{bmatrix}\Sigma\begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix}$, $H_e=Z^TY^TH_0YZ$,

$E_1=\begin{bmatrix} B_0 & B_1 & B_2N \end{bmatrix}\begin{bmatrix} c^T & \tilde{x}^T & \tilde{d}^T \end{bmatrix}^T$ and $E_2=\begin{bmatrix} B_1 & B_2N \end{bmatrix}\begin{bmatrix} a^T & \delta^T \end{bmatrix}^T$.

Table B.3 Statistical Tests

Statistical Test	Based on
z_e	$e \sim N(0, H_e)$
z_r	$r \sim N(0, H_r)$
$z_{a,\delta}$	$\begin{bmatrix} a \\ \delta \end{bmatrix} \sim N(0, Q)$
z_e^*	$\lambda \sim N\left(0, H_e^{-1}\right)$
z_r^*	$\mu \sim N\left(0, YZH_e^{-1}Z^TY^T\right)$
$z_{a,\delta}^*$	$\Sigma^{-1}\begin{bmatrix} a \\ \delta \end{bmatrix} \sim N\left(0, \Sigma^{-1}Q\Sigma^{-1}\right)$ or $\Sigma^{-1}\begin{bmatrix} a \\ \delta \end{bmatrix} \sim N\left(0, \begin{bmatrix} B_1^T \\ NB_2^T \end{bmatrix} YZH_e^{-1}Z^TY^T \begin{bmatrix} B_1 & B_2N \end{bmatrix}\right)$
y_e, Q_e, and $\chi^2_{k_e}$	$e \sim N(0, H_e)$ and eigensystem of H_e
y_r, Q_r, and $\chi^2_{k_r}$	$r \sim N(0, H_r)$ and eigensystem of H_r
$y_{a,\delta}$, $Q_{a,\delta}$, and $\chi^2_{k_{a,\delta}}$	$\begin{bmatrix} a \\ \delta \end{bmatrix} \sim N(0, Q)$ and eigensystem of Q

Literature Cited

Albers, J. E., "Data Reconciliation with Unmeasured Variables," *Hydrocarbon Processing*, **73**, 65, (1994).

Aldrich, C., and J. S. J. Vandeventer, "Identification of Gross Errors in Material Balance Measurements by Means of Neural Nets," *Chemical Engineering Science*, **49**, 1357, (1994).

Almasy, G. A., "Principles of Dynamic Balancing," *AIChE J.* **36**, 1321, (1990).

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

Bailey, J. K., "Nonlinear Optimization of a Hydrocraker Fractionation Plant," *M. Eng. Thesis*, McMaster University, Hamilton, Ontario, Canada, (1991).

Bossen, B. S., L. J. Christiansen, J. E. Jarvan, and R. Gani, "Simulation, Optimization and Data-Reconciliation of Industrial-Chemical Processes," *Chemical Engineering Research & Design*, 72, 376, (1994).

Crowe, C. M., "Formulation of Linear Data Reconciliation Using Information Theory," submitted to *Chemical Engineering Science*, (1995).

Crowe, C. M., "Data Reconciliation - Progress and Challenges," *Proceedings of the 5th International Symposium on Process Systems Engineering*, Kyongju, Korea, **1**, 111, (May, 1994).

Crowe, C. M., "The Maximum-Power Test for Gross Errors in the Original Constraints in Data Reconciliation," *Can. J. of Chem. Eng.*, **70**, 1030, (1992).

Crowe, C. M., "Observability and Redundancy of Process Data for Steady State Reconciliation," *Chemical Engineering Science*, **44**, 2909, (1989a).

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

Crowe, C. M., "Recursive Identification of Gross Errors in Linear Data Reconciliation," *AIChE J.*, **34**, 541, (1988).

Crowe, C. M., "Reconciliation of Process Flow Rates by Matrix Projection. II. The Nonlinear Case," *AIChE J.*, **32**, 616, (1986).

Crowe, C. M., Y. A. Garcia Campos and A. Hrymak, "Reconciliation of Process Flow Rates by Matrix Projection. I. The Linear Case," *AIChE J.*, **29**, 818, (1983).

Darouach M., and M. Zasadzinski, "Data Reconciliation in Generalized Linear Dynamic Systems," *AIChE J.*, **37**, 193, (1991).

Hodouin, D., and M. D. Everell, "A Hierarchical Procedure for Adjustment and Material Balancing of Mineral Processes Data," *International Journal of Mineral Processing*, 7, 91, (1980).

Holly, W., R. Cook, and C. M. Crowe, "Reconciliation of Mass Flow Rate Measurements in a Chemical Extraction Plant," *Can. J. of Chem. Eng.*, **67**, 595, (1989).

Horn, J. L., "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, **30**, 179, (1965).

Hotelling, H., "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educ. Psychol.*, **24**, 417, (1933).

Islam, K. A., G. H. Weiss, and J. A. Romagnoli, "Nonlinear Data Reconciliation for an Industrial Pyrolysis Reactor," *Computers & Chemical Engineering*, **18**, S217, (1994).

Jackson, J. E., and R. A. Bradley, "Sequential χ^2 and T^2-tests," *Annals of Mathematical Statistics*, **32**, 1063, (1961a).

Jackson, J. E., and R. A. Bradley, "Sequential χ^2- and T^2-tests and Their Application to an Acceptance Sampling Problem," *Technometrics*, **3**, 519, (1961b).

Jackson, J. E., "A User's Guide to Principal Components," *John Wiley & Sons, Inc.*, New York, (1991).

Jolliffe, I. T., "Principal Component Analysis," *Springer-Verlag*, New York, (1986).

Kao, C. S., A. C. Tamhane, and R. S. H. Mah, "Gross Error Detection in Serially Correlated Process Data," *Industrial & Engineering Chemistry Research*, **29**, 1004, (1990).

Klemsa, J., and T. Perris, "Quality Information for Production Management," *Computers & Chemical Engineering*, **16**, S507, (1992).

Kresta, J., J. F. MacGregor, and T. E. Marlin, "Multivariate Statistical Monitoring of Process Operating Performance," *Can. J. of Chem. Eng.*, **69**, 35, (1991).

Kuehn, D. R., and H. Davidson, "Computer Control. II Mathematics of Control," *Chem. Eng. Prog.*, **57**, 44, (1961).

Kim, I. W., M. J. Liebman, and T. F. Edgar, "Robust Error-in-Variables Estimation Using Nonlinear Programming Techniques," *AIChE J.*, **36**, 985, (1990).

Liebman, M. J., T. F. Edgar, L. S. Lasdon, "Efficient Data Reconciliation and Estimation for Dynamic Processes Using Nonlinear Programming Techniques," *Computers & Chemical Engineering*, **16**, 963,(1992).

Lawrence, R. J., "Data Reconciliation: Getting Better Information," *Hydrocarbon Processing*, 55, (June, 1989).

Liu, Y., and D. S. Blostein, "Optimality of the Sequential Probability Ratio Test for Nonstationary Observations," *IEEE Transactions on Information Theory*, **38**, 177, (1992).

MacGregor, J. F., C. Jaeckle, C. Kiparissides, and M. Koutoudi, "Monitoring and Diagnosis of Process Operating Performance by Multi-Block PLS Methods with an Application to Low Density Polyethylene Production," *AIChE J.* **40**, 826, (1994).

Madron, F., "Process Plant Performance. Measurement and Data Processing for Optimization and Retrofits," *Ellis Horwood Ltd.*, Chichester, England, (1992).

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

Mah, R. S. H., "Data Screening," *Foundations of Computer-Aided Process Operations*, G. V. Reklaitis, H. D. Spriggs, eds., Elsevier, 67, (1987).

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

"Manual of Mathematics", *in Chinese*, 618, Beijing, China, (1979).

Maquin, D., M. Darouach, and J. Ragot, "Observability and Data Validation of Bilinear Constraints," *IFAC-AIPAC'89 Symposium*, Nancy, France, (July, 1989).

Markos, J., and M. Barto, "Optimization Method for Balance Adjustment In Large Chemical Process Systems," *Chemical Engineering and Processing*, **30**, 45, (1991).

Narasimhan S., and P. Harikumar, "Method to Incorporate Bounds in Data Reconciliation and Gross Error Detection - I. The Bounded Data Reconciliation Problem," *Computers & Chemical Engineering*, **17**, 1115, (1993).

Narasimhan S., R. S. H. Mah, A. C. Tamhane, J. W. Woodward, and J. C. Hale, "A Composite Statistical Test for Detecting Changes of Steady States," *AIChE J.*, **32**, 1409, (1986).

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

Pages, A., H. Pingaud, M. Meyer, and X. Joulia, "A Strategy for Simultaneous Data Reconciliation and Parameter-Estimation on Process Flowsheets," *Computers & Chemical Engineering*, **18**, S223, (1994).

Pearson, K., "On Lines and Planes of Closest Fit to Systems of Points in Space," *Phil. Mag., Ser. B*, **2**, 559, (1901).

Ragot, J., D. Maquin, and D. Sauter, "Data Validation Using Orthogonal Filters," *IEE Proceedings-D*, **139**, 47, (1992).

Reilly, P. M., and R. E. Carpani, "Application of Statistical Theory of Adjustments to Material Balances," *13th Can. Chem. Eng. Conf.*, Montreal, Oue., (Oct., 1963).

Rushton, S., "On a Sequential t-Test," *Biometrica*, **37**, 326, (1950).

Rushton, S., "On a Two-Sided Sequential t-Test," *Biometrica*, **39**, 302, (1952).

Sanchez, M., A. Bandoni, and J. Romagnoli, "PLADAT - A Package for Process Variable Classification and Plant-Data Reconciliation," *Computers & Chemical Engineering*, **16**, S499, (1992).

Seber, G. A. F., "Multivariate Observations," *John Wiley and Sons, Inc.*, New York, (1984).

Schraa, O., "The Numerical Solution of Bilinear Data Reconciliation Problems Using Unconstrained Optimization Methods," *M. Eng. Thesis*, McMaster University, Hamilton, Ontario, Canada, (1995).

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

Siegmund, D., "Sequential Analysis — Tests and Confidence Intervals," *Springer-Verlag*, New York, (1985).

Smith, H. W., and N. Ichiyen, "Computer Adjustment of Metallurgical Balances," *CIM Bulletin*, **66**, 97, (Sept., 1973).

Stanley, G. M., and R. S. H. Mah, "Observability and Redundancy in Process Data Estimation," *Chemical Engineering Science*, **36**, 259, (1981).

Tjoa, I. B., and L. T. Biegler, "Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems," *Computers & Chemical Engineering*, **15**, 679, (1991).

Tong, H., and C. M. Crowe, "Detection of Gross Errors in Data Reconciliation by Principal Component Analysis," *AICHE J.*, **41**, 1712, (1995).

Tong, H., and C. M. Crowe, "Principal Component Test in On-Line Data Reconciliation," *43rd Can. Chem. Eng. Conf.*, Ottawa, Ont., (Oct., 1993).

Van der Heijden, R. T. J. M., J. J. Heijnen, C. Hellinga, B. Romein, and K. Ch. A. M. Luyben, "Linear Constraint Relations in Biochemical Reaction Systems: I. Classification of the Calculability and the Balanceability of Conversion Rates," *Biotechnology and Bioengineering*, **43**, 3, (1994a).

Van der Heijden, R. T. J. M., B. Romein, J. J. Heijnen, C. Hellinga, and K. Ch. A. M. Luyben, "Linear Constraint Relations in Biochemical Reaction Systems: II. Diagnosis and Estimation of Gross Errors," *Biotechnology and Bioengineering*, **43**, 11, (1994b).

Veverka, V., "A Method of Reconciliation of Measured Data with Nonlinear Constraints," *Applied Mathematics and Computation*, **49**, 141, (1992).

Wald, A., "Sequential Analysis," *John Wiley & Sons, Inc.*, New York, (1947).

Whittle, P., "Optimization Over Time," Vol. I and II, *John Wiley & Sons, Inc.*, New York, (1982, 1983).

Wold, S., K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory systems*, **2**, 37, (1987).