Detecting Persistent Gross Errors by Sequential Analysis of Principal Components

Hongwei Tong and Cameron M. Crowe
Dept. of Chemical Engineering, McMaster University, Hamilton, Ont., Canada L8S 4L7

Measurements such as flow rates from a chemical process violate conservation laws and other process constraints because they are contaminated by random errors and possibly gross errors such as process disturbances, leaks, departures from steady state, and biased instrumentation. Data reconciliation is aimed at estimating the true values of measured variables that are consistent with the constraints, at detecting gross errors, and at solving for unmeasured variables. An approach to constructing sequential principal-component tests for detecting and identifying persistent gross errors during data reconciliation by combining principal-component analysis and sequential analysis is presented. The tests detect gross errors as early as possible with fewer measurements. They were sharper in detecting and have a substantially greater power in correctly identifying gross errors than the currently used statistical tests in data reconciliation.

Introduction

Data obtained from a chemical process are inherently inaccurate and thus violate conservation laws and other process constraints. The purpose of data reconciliation is to resolve the contradictions between the measurements and their constraints, and to process contaminated data into consistent information. Since meaningful data adjustments can be obtained if and only if there is no gross error in the data, gross errors must be detected and then eliminated or corrected before a valid data reconciliation can be achieved.

Conventional tests for gross error, such as the chi-square test on the final minimum weighted sum of squares of adjustments to the measurements, or tests on the residuals of constraints or on individual adjustments have not been entirely reliable in correctly identifying the gross errors. This has arisen because of correlation among measurements and among the resulting adjustments and residuals. This work is aimed at detecting persistent gross errors by combining the previously reported principal component (PC) test with the sequential analysis approach of Wald (1947). The tests are designed to detect gross errors as early as possible with fewer measurements than other tests and to reduce the number of erroneous conclusions about the presence or absence of gross errors and their identification.

A transient gross error, caused by process disturbances, even if it is relatively large, occurs randomly for an instant and neither persists nor does it affect process performance. However, a persistent gross error, even if it is relatively small, is deterministic and usually results from sensor problems or process leaks that potentially seriously affect the process. When persistent gross errors occur, we should detect them as early as possible in order to limit their influence on process operation and product quality, and to prevent them from contaminating plant databases.

Steady-State Linear Data Reconciliation

Steady-state linear data reconciliation is defined as a quadratic programming problem (Crowe et al., 1983; Tong and Crowe, 1995; Crowe, 1996)

$$\mathcal{P}: \quad \min_{a} F(a) = a^T \Sigma_1^{-1} a \tag{1}$$

$$\text{s.t. } B_1(\tilde{x} + a) + Pv = \mathbf{0}, \tag{2}$$

where the data adjustments a are minimized subject to process constraints (Eq. 2). In Problem $\mathcal{P}$, $\tilde{x}$ is the vector of measured data (for instance, $\tilde{x}$ consists of component flow rates if material balances are considered); v is the vector of un-

Correspondence concerning this article should be addressed to Cameron M. Crowe.
Present address of Hongwei Tong: Simulation Sciences Inc., 601 Valencia Ave., Suite 100, Brea, CA 92823.

measured variables such as unmeasured component flow rates or extents of reactions; B_1 and P are the matrices of coefficients corresponding to $\tilde{x}$ and v in Eq. 2; Σ_1 is the variance–covariance matrix of $\tilde{x}$. The process constraints expressed in Eq. 2 are called the *original process constraints.*

The unmeasured variables can be removed by using matrix projection (Crowe et al., 1983) such that the original constraints in Problem $\mathcal{P}$ are transformed into a set of *reduced process constraints* that retain only the redundant measured flow rates. The set of residual balances is not unique, but the optimal solution is unique. However, the result of a conventional test of a residual constraint for gross error will depend on the set of constraints retained. The variance–covariance matrix of the reduced constraints is

$$H_e = \text{cov}(e) = B_1^R \Sigma_1 (B_1^R)^T, \tag{3}$$

where B_1^R is the balance matrix of the reduced constraints, whose residual vector is $e = B_1^R \tilde{x}$.

It was shown that the optimal value of the objective function is

$$F = \chi_m^2 = e^T H_e^{-1} e, \tag{4}$$

which follows a chi-square distribution with m degrees of freedom (Crowe et al., 1983), where m is the rank of H_e.

The process constraints given in Problem $\mathcal{P}$ do not impose a restriction that $(\tilde{x} + a)$ cannot be negative, so it is possible to get an infeasible adjustment to a measured component flow rate. Nonnegativity constraints can be easily implemented in a general nonlinear programming framework, so a reconciled flow that would otherwise be negative is set to zero. This result is still likely to be caused by gross errors. When such errors are identified and removed, feasible and optimal adjustments should be obtained in most cases, without imposing any nonnegativity constraint. Since the presence of an infeasible solution is itself an indication that a thorough inspection of the process and the instrumentation should be made, nonnegativity constraints have not been included in the reconciliation model.

Crowe (1986) and Tong (1995) studied the problem of bilinear data reconciliation. The statistical tests that we present can also be applied to the bilinear problem.

Statistical Tests for Gross Errors

Two important objectives of data reconciliation are optimally to adjust measurements and detect gross errors. The former is an optimization problem, and the latter is a statistical hypothesis testing problem. The final optimal adjustments to measurements, which are consistent with the null hypothesis that the true values of the adjustments are zero, should be made only when the measurements in gross error have been removed or corrected.

There are two fundamental steps in gross-error correction, namely showing that at least one gross error is present (detection) and then identifying which measurement(s) and/or constraint(s) contain the gross error(s) (identification).

A gross error was described by Van der Heijden et al. (1994) as one that is relatively large compared to the variable's variance. This is not wholly acceptable because a gross error in a variable can be relatively small compared to its variance. It would be more appropriate to state that a gross error in a measurement or a constraint violates the process covariance structure.

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 control objectives or quality specifications. Once a gross error is detected, it may also be difficult to locate it because the correlation caused by the process topology confounds the statistics. Confounding statistics may be misleading and may result in throwing away good data while keeping corrupted ones. Tong and Crowe (1995) presented a method based on principal component analysis (PCA) that appropriately handles the correlation among the variables.

Currently Used Statistical Tests for Gross Errors

Several statistical tests have been developed to detect gross errors. Tong and Crowe (1995) reviewed the chi-square collective test, which compares the optimal value of the objective function in the model of data reconciliation to an appropriate tabulated chi-square value, as well as the univariate and the maximum power (MP) constraint and measurement tests that examine each residual of the process constraints and each measurement adjustment.

Though those statistical tests have been widely used in detecting gross errors, their performance has not always been satisfactory. The problem is caused by ignoring variable correlations (the univariate tests) and by limited accounting for those correlations (the MP tests). The correlations are inherent not only because of certain common factors such as a total flow rate shared by several component flows in a stream, but also because of the process topology. The correlations usually confound the test statistics and make it very difficult to identify gross errors (Tong and Crowe, 1995). When the tests fail to detect gross errors, the interpretation of plant performance may be distorted.

Principal-Component Tests for Gross Errors

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. Tong and Crowe (1995) derived the PC tests aimed at overcoming problems with the previous tests.

Principal-component transformation

Consider a set of linear combinations of the reduced balance residuals, e

$$y_e = W_e^T e = \Lambda_e^{-1/2} U_e^T e. \tag{5}$$

Matrix Λ_e is diagonal, consisting of the eigenvalues $\lambda_{e,i}$, $i = 1, \ldots, m$, of H_e, the variance–covariance matrix of e. Matrix U_e consists of the orthonormalized eigenvectors of H_e, so that

$$U_e U_e^T = I. \tag{6}$$

Vector y_e contains the principal components, with the values of its elements being called the principal-component scores.

It can be shown that $y_e \sim (\mathbf{0}, I)$ because $e \sim (\mathbf{0}, H_e)$, so that a set of correlated variables, e, is transformed into a new set of uncorrelated variables, y_e, with unit variances. The elements (PCs) of y_e are arranged to correspond to the eigenvalues of H_e listed in descending order of magnitude. In particular, if the measured variables are normally distributed about mean values, x, that obey the reduced balances, namely $\tilde{x} \sim N(x, \Sigma_1)$, then

$$y_e \sim N(\mathbf{0}, I). \tag{7}$$

Test of a principal component

Based on Eqs. 5 and 7, the test statistic for a principal component is defined as

$$y_{e,i} = (W_e^T e)_i \sim N(0,1), \qquad i = 1, \ldots, m, \tag{8}$$

which can be tested against a threshold tabulated value. Equation 8 shows that the ith principal component, $y_{e,i}$, is obtained from the inner product of e and the ith eigenvector from W_e.

Contribution analysis

Once a gross error is detected, it has to be identified. A diagnostic method for finding the causes of outliers by interrogating a PCA model was discussed by MacGregor et al. (1994) and Tong and Crowe (1995). The constraints in gross error can be identified by inspecting the contribution from the jth residual, e_j, to a suspect principal component, say $y_{e,i}$, for $j = 1, 2, \ldots$. The first few constraints, which make the larger contributions, are called the *major contributors,* and are considered to be in gross error. Tong and Crowe (1995) also discussed how to obtain the *exact* probability of a type I error for a PC from a prescribed overall type I error.

PC tests of the original constraints and the adjustments

Tong and Crowe (1995) showed that the PC tests could also be applied to the original constraints and to the adjustments to the measurements. Additional detail can be found in Tong (1995).

Transient and Persistent Gross Errors

It is important to know if a gross error is transient or persistent, and what the probabilities are of being wrong in detecting or failing to detect such an error. The statistical tests given earlier cannot efficiently distinguish a persistent gross error from a transient one, if they test one set of measurements at a time, without taking into account any other measurements sampled immediately before them.

Nomikos and MacGregor (1994) used PCA to monitor a batch process. They plot the first few PC scores, one vs. another, and check whether there is any cluster away from the origin. Such a cluster proves that the process is shifted away from the normal operating region. They also plot the loadings to detect the variables responsible for the abnormal operation. In fact, their technology could be used in data reconciliation to detect and identify gross errors. Specifically, score plots could be used in gross-error detection, and loading plots could be used in gross-error identification.

However, there are differences between gross-error detection in data reconciliation and in process monitoring. Among these differences are

1. Use of the constraint residuals or data adjustments vs. the original data
2. Use of the topology built into the covariance matrix vs. no use of process topology
3. Retention of all the PCs vs. retention of only the first few PCs.

As we already discussed, gross-error detection requires the analysis of all the PCs, which makes graphing cumbersome. Furthermore, in order to carry out cluster analysis, more data must be sampled than are needed for data reconciliation.

Sequential Analysis

Sequential analysis was pioneered by Wald (1947) and advanced by other researchers (Jackson and Bradley, 1961a,b; 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% in the number of observations over the most efficient nonsequential 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). A sequential principal-component test is presented that takes advantage of both the principal-component tests and sequential analysis.

Hypotheses

Any of the principal components derived from e is distributed with the zero mean and the unit variance if the measurements are free of gross errors. A principal component is normally distributed if the measurements are free of gross errors.

It is natural therefore to test whether the mean of a principal component is zero. Let y be any 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 the possibility of having gross errors. One is 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 δ such that the acceptance of the null hypothesis is regarded as an error of practical importance whenever

$$|\theta| \geq \delta \tag{9}$$

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

Sequential Sampling Scheme with Prescribed α and β. A sampling scheme is needed for which the probability that the null hypothesis will be rejected does not exceed a small prescribed value α whenever $\theta = 0$ (type I error), and the probability of accepting the null hypothesis does not exceed a small prescribed value β whenever the alternative hypothesis is true (type II error). It is impossible to estimate the probability of type II error in the hypothesis testing because the fault size is unknown. A particular method of sequential analysis, known as the *sequential probability ratio test* (SPRT), was developed by Wald (1947). The method can be adapted to test a principal component for persistent gross errors. The method can of course be applied directly to $\boldsymbol{e}$, $\boldsymbol{r}$, and $\boldsymbol{a}$. However, tests for principal components usually result in sharper detection and less confounding identification.

Sequential probability ratio test

Let $f(y, \theta)$ denote the probability density function of a principal component. Let H_0 be the null hypothesis that $\theta = 0$, 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. The successive observations on y are denoted by y^1, y^2, ..., and so on.

Let $p_{ij} = p_{ij}(y^1, \ldots, y^j)$ denote the joint probability density function that a sample y^1, ..., y^j is obtained for any integer $j > 0$ when H_i is true, for $i = 0$ or 1. Of course if the successive observations y^1, y^2, ..., are independent observations on y, $p_{1j} = f(y^1, \theta) \cdots f(y^j, \theta)$ 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 p_{1j}/p_{0j} is computed and compared to two constants A and B $(0 < B < A)$, as illustrated below:

$p_{1j}/p_{0j} \leq B$	$B < p_{1j}/p_{0j} < A$	$p_{1j}/p_{0j} \geq A$
Accept H_0 (no gross error) $_B$	Take one more observation $_A$	Reject H_0 (gross error)

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

$$A \leq (1 - \beta)/\alpha \tag{10}$$

$$B \geq \beta/(1 - \alpha), \tag{11}$$

and for all practical purposes equality can be chosen as the criterion. This results in a slight increase in the number of observations, which would be acceptable for a flow rate and temperature measurements that are not excessively costly.

Since the magnitude of any gross error is unknown in advance, the true probability β of a type II error is unknown. The chosen value of β represents an upper bound on the true value corresponding to the limiting case of a gross error with magnitude δ.

It can be shown that if the successive observations y^1, y^2, ... are independent observations on y, the sequential process will eventually terminate with the probability of one. However, requiring the assumption of independent sampling 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, ..., y^{j-1}, which makes the successive observations dependent. Fortunately, the inequalities 10 and 11 remain valid in spite of the dependence of the successive observations, provided that the procedure will eventually terminate (Wald, 1947, p. 43). The method is valid in general for dependent observations.

Collecting Data for Sequential Analysis. The reason for 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 smaller than that of 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.

Univariate Sequential Test of a PC. An adequate sampling scheme for testing that the mean of a principal component is zero is given as follows. Following Wald (1947, p. 84), the ratio

$$\frac{p_{1j}}{p_{0j}} = \frac{1}{2}\,\frac{\exp\left[-\frac{1}{2}\sum_{i=1}^{j}(y^i - \delta)^2\right] + \exp\left[-\frac{1}{2}\sum_{i=1}^{j}(y^i + \delta)^2\right]}{\exp\left[-\frac{1}{2}\sum_{i=1}^{j}(y^i)^2\right]}$$

$$= \exp\left(-\frac{j\delta^2}{2}\right)\cosh\left(\delta \sum_{i=1}^{j} y^i\right) \tag{12}$$

is computed as each set of measurements is available. (We can only measure flow rates and concentrations, not principal components. However, for simplicity, we will speak in terms

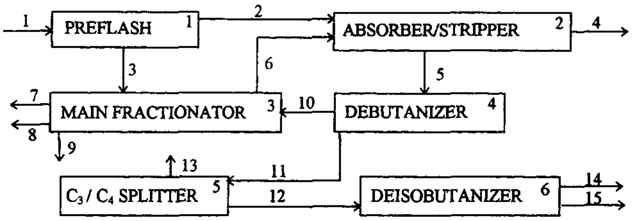

Figure 1. Sunoco hydrocracker fractionation plant.

of the measurements of principal components. This should not cause any confusion.) The computation will continue as long as $B < p_{1j}/p_{0j} < A$. The null hypothesis that the principal component is not in gross error is accepted if $p_{1j}/p_{0j} \leq B$, and rejected if $p_{1j}/p_{0j} \geq A$, where A and B are given by Eqs. 10 and 11. In fact, the tests were done on $Z_j = \log\cosh(\delta \sum_{i=1}^{j} y^i) = \log(p_{1j}/p_{0j}) + j\delta^2/2$.

Guidelines for choosing α, β, and δ are given in the following example.

Identifying a Gross Error

After a gross error has been detected by the univariate sequential test for the principal components, it can be identified by the method of contribution analysis discussed in Tong and Crowe (1995). 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.

Example: Hydrocracker fractionation plant

The hydrocracker fractionation plant, studied by Bailey (1991) and shown in Figure 1, includes 6 process units and 15 streams. There is assumed to be no covariance among the measurements of the total mass flow rates to be reconciled.

Choosing 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 nonsequential tests. An inappropriate choice may result in more samples being taken or a wrong inference. The following guidelines are recommended for choosing a consistent set of parameters α, β, and δ.

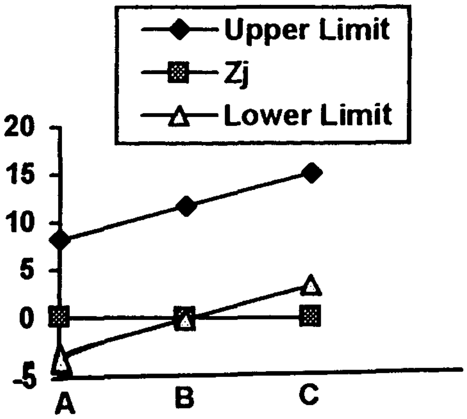

Figure 2. Sequential test for $y_{a,1}$.

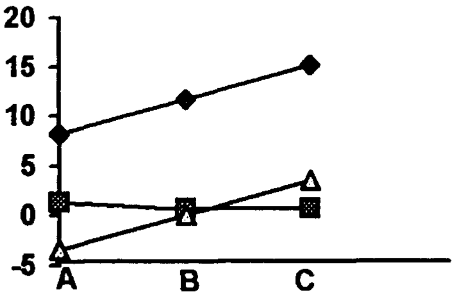

Figure 3. Sequential test for $y_{a,2}$.

1. α should be obtained from the prescribed probability of the overall type I error. In this example, the desired probability of the overall type I error is specified as 0.05. With six degrees of freedom (six principal components), $\alpha = 0.0085$ with the threshold of 2.631.

2. β should be small enough to keep the test sharp, because β is the probability of the error that one could commit if a principal component is subject to persist gross errors while one infers otherwise. A gross error can significantly distort data reconciliation and usually relates to sensor failure or a process leak. The value $\beta = 0.001$ is chosen in this example.

3. The choice of δ reflects one's judgment as to how far away from zero is the absolute value of the mean of a principal component considered to be evidence of gross error. Since a principal component 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, $\delta = 2.6$.

Sequential Analysis for the Principal Components. The data obtained at 7 A.M. and 7 P.M. on February 26, 1986 (Day 1), February 27, 1987 (Day 2), and March 3, 1987 (Day 3) are given in Tong (1995). The morning sets of measurements are labeled A, B, C and the evening sets are D, E, and F, in respective chronological order.

In the first study, we look at the data sorted with the morning data followed by the evening data. This arrangement would reveal any shift to shift difference. The results of sequential analysis are given in Figures 2 to 7. For illustration purposes, Figures 2 to 5 only show a single round of sequential test—the first time when a sequential test accepts H_0.

It can be seen from Figures 2 through 5 that principal components 1 through 4 are not subject to any gross error, because Z_j eventually went below the lower limit in each case.

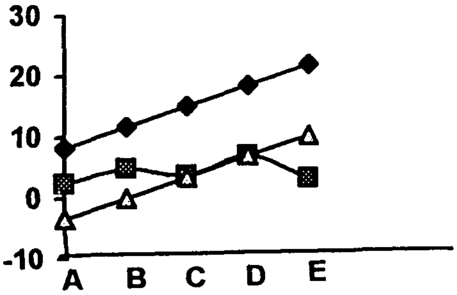

Figure 4. Sequential test for $y_{a,3}$.

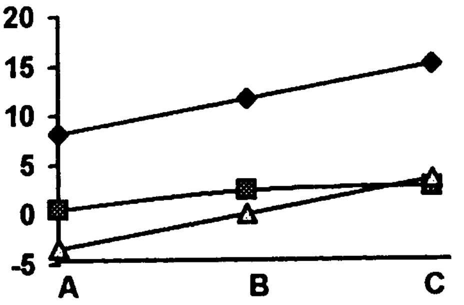

Figure 5. Sequential test for $y_{a,4}$.

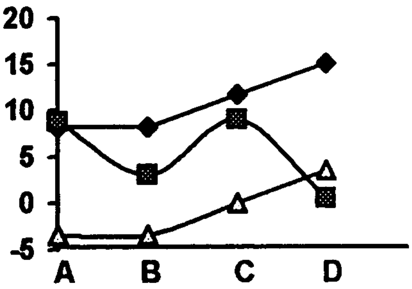

Figure 7. Sequential test for $y_{a,6}$.

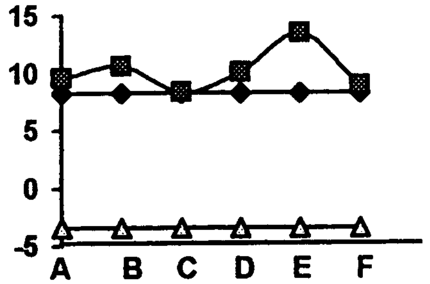

Figure 6. Sequential test for $y_{a,5}$.

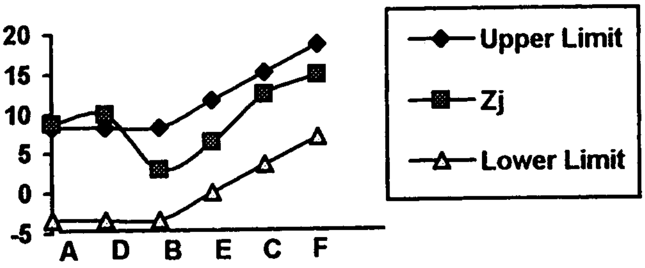

Figure 8. Another sequential test for $y_{a,6}$.

Principal components 5 and 6 are subject to gross error. With a closer look at Figures 6 and 7, one finds that the statistic for principal component 5 is always above the upper limit, which indicates that the principal component is subject to a persistent gross error. However, the statistic for principal component 6 is above the upper limit in test set A, but below the lower limit at D in the second group (consisting of sets B, C, and D). In the third group (consisting of sets E and F, not shown in Figure 7), the statistic is between the upper and lower limits. One should note that the change in slope of the threshold levels in Figure 7 results from having restarted the sequential analysis after the first test set A. The result shows no evidence of shift-to-shift difference. However, is principal component 6 also subject to persistent gross error?

This question is investigated by performing another study, where the data are arranged chronologically, that is, in the sequence of A, D, B, E, C, F. The result is given in Figure 8. In the first two test sets (A and D) the sequential test statistics corresponding to the data sampled on Day 1 are above the upper limit, while for the next group (consisting of sets 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. Again one should note that the sequential analysis is restarted as soon as Z_j goes above the upper limit. It is restarted three times in Figure 8 at A, D, and B. That also explains the difference in slope of the threshold levels between Figures 7 and 8. In Figure 7, the value of Z_j at D depends upon B, C, and D, since these sets are in the same group. In Figure 8, however, the value of Z_j at D is independent, because D is the only set in that group. Therefore, Z_j at D has a different value in Figure 7 (morning–evening order) than that in Figure 8 (chronological order). This suggests that principal component 6 was likely in 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

Table 1. Hydrocracker Fractionation Plant

	$\tilde{x}$	Variance	$\hat{x}$	$z_a = z_a^*$	y_a	Contrib. to $y_{a,5}$	Contrib. to $y_{a,6}$
1	4,769.20	56,863	4,887.89	0.559	−0.266	0.004	0.009
2	356.02	316.9	424.42	**5.346**	0.750	**−1.759**	**3.601**
3	4,391.4	48,211	4,463.47	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	1,005.9	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	2,191.2	12,004	2,148.2	−0.782		0.001	0.001
9	1,909.40	9,114.4	1,876.75	−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

fact, it would be indifferent, based on the information available to infer whether that principal component was in gross error. Our best guess is that it might be in persistent gross error considering the trend of the Z_j in Figure 8 and the fact that the data sets were separated by a period of one year. In the plant, it would be helpful to check the sensors and the process units related to that principal component. Moreover, more measurements might have provided a better inference.

Identifying the Gross Error. Once it is known that the data were subject to persistent gross error, the contribution analysis given in Tong and Crowe (1995) and Tong (1995) can be used to identify the error.

As an example, the data taken at 7:00 A.M., Day 1, are studied. Table 1 shows the contributions from the measurement adjustments to the outlier PCs, with the major contributors ($\mathbf{a}_2$, $\mathbf{a}_{12}$) being in boldface. It also shows the reconciled measurements, univariate, and PC tests of the measurement adjustments (z_a and y_a). There is no covariance among the measurements in this case. The threshold for a uninormal variate with 6 degrees of freedom is 2.631, and that for a chi-square variate with the same degrees of freedom is 12.59.

Table 2 shows what would happen if the suspect measurements flagged by the different statistics were deleted one at a time. Table 3 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 outlier is deleted.

PC tests of $\mathbf{a}$ 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, since if $\tilde{x}_4$ were deleted, the reconciliation would be infeasible because $\hat{x}_4$ becomes negative. If another suspect $\tilde{x}_6$ given by $z_{a,6}$ 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.

In this example, the PC tests are seen to be more effective than the other tests for detection and identification of gross errors. Table 4 summarizes the results that all statistical tests were passed when the suspect measurement $\tilde{x}_2$ was deleted.

Table 2. Deletion of a Variable

Variable Deleted	Result	Comment
$\tilde{x}_2$	All tests passed	$\chi_m^2 = 1.73$, **lowest**
$\tilde{x}_{12}$	*Some outliers*	
$\tilde{x}_4$	$\hat{x}_4$ *negative*	*Infeasible*
$\tilde{x}_6$	All tests passed	$\chi_m^2 = 2.19$, higher
$\tilde{x}_{10}$	*Some outliers*	
$\tilde{x}_{13}$	*Some outliers*	

Table 4. $\tilde{x}_2$ Deleted

	$\tilde{x}$	Variance	$\hat{x}$	$z_a = z_a^*$	y_a
1	4,769.20	56,863	4,914.17	0.683	−0.451
2	—	—	(488.43)	—	0.649
3	4,391.4	48,211	4,425.75	0.180	0.219
4	62.06	9.63	62.04	−0.683	0.608
5	634.31	1,005.9	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	2,191.2	12,004	2,152.04	−0.712	
9	1,909.40	9,114.4	1,879.67	−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	

Conclusions and Significance

The solution for the model of steady-state data reconciliation provides the residuals $\mathbf{r}$ of the original and $\mathbf{e}$ of the reduced constraints and the adjustments $\mathbf{a}$, together with their variance–covariance matrices. Their principal components are obtained by performing the principal-component transformation. They are mutually independent, and indeed, can be tested one by one. Each principal component is then subject to the sequential analysis for detecting any persistent gross errors. Once a persistent gross error is detected, the contribution analysis (Tong and Crowe, 1995) is carried out to identify the major contributors to the suspect PCs. These major contributors are the gross errors.

The model of steady-state data reconciliation gives *a priori* knowledge that the expectations of $\mathbf{e}$, $\mathbf{r}$, and $\mathbf{a}$ are zero in the absence of gross error. The PCA on $\mathbf{e}$, $\mathbf{r}$, and $\mathbf{a}$ removes the correlation and transfers them into uncorrelated PCs. The principal-component tests derived from such an analysis have been found to be sharper in detecting and less confounding in identifying gross errors than the other currently used statistical tests in data reconciliation. A sequential test procedure was then developed by applying sequential analysis to the PCs.

The number of measurements required in the 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 for detecting persistent gross errors. Desired probabilities of type I and II errors that one may commit in detecting gross errors can be specified in advance by adjusting the parameters in the test.

Table 3. Tests of $\mathbf{a}$

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
$y_{a,6}$	$\tilde{x}_2$	The primary suspect $\tilde{x}_2$ is correct

Literature Cited

Bailey, J. K., "Nonlinear Optimization of a Hydrocracker Fractionation Plant," M. Eng. Thesis, McMaster Univ., Hamilton, Ont., Canada (1991).

Crowe, C. M., "Data Reconciliation—Progress and Challenges," *J. Proc. Cont.*, **6**, 89 (1996).

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).

Jackson, J. E., and R. A. Bradley, "Sequential χ^2 and T^2-tests," *Ann. Math. Stat.*, **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).

Liu, Y., and D. S. Blostein, "Optimality of the Sequential Probability Ratio Test for Nonstationary Observations," *IEEE Trans. Inform. Theory*, **IT-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).

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

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

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

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

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

Tong, H., "Studies in Data Reconciliation Using Principal Component Analysis," PhD Thesis, McMaster Univ., Hamilton, Ont., Canada (1995).

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," *Biotechnol. Bioeng.*, **43**, 11 (1994).

Wald, A., *Sequential Analysis*, Wiley, New York (1947).

Whittle, P., *Optimization Over Time*, Vols. I and II, Wiley, New York (1982, 1983).

Manuscript received Apr. 24, 1996, and revision received Dec. 23, 1996.