Detection of Gross Errors in Data Reconciliation by Principal Component Analysis

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

Statistical testing provides a tool for engineers and operators to judge the validity of process measurements and data reconciliation. Univariate, maximum power and chi-square tests have been widely used for this purpose. Their performance, however, has not always been satisfactory. A new class of test statistics for detection and identification of gross errors is presented based on principal component analysis and is compared to the other statistics. It is shown that the new test is capable of detecting gross errors of small magnitudes and has substantial power to correctly identify the variables in error, when the other tests fail.

Introduction

Measured process data are inherently inaccurate and violate process constraints because of their underlying stochastic properties and possible gross errors caused by process disturbances, process leaks, malfunctioning or miscalibrated instrumentation, or even departure from steady state. The theory of data reconciliation has been developed to resolve the contradictions between the measurements and their constraints and to process contaminated data into consistent information. Reviews have been published by Crowe et al. (1983), Crowe (1986, 1994), Mah (1987), and Ragot et al. (1992). A good introduction can be found in Lawrence (1989).

Several statistical tests have been defined to detect 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 mathematical 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, and is called the MP measurement test. The MP constraint tests were proposed by Crowe (1989, 1992).

Although those tests have been widely used in detecting gross errors, their performance has not always been satisfactory, particularly when the gross errors are subtle. When the tests fail to detect the errors, the interpretation of plant performance is distorted. In some cases, the tests will detect the presence of gross errors but will not correctly identify the variable in error. Therefore, one may risk deleting good measurements and keeping the corrupted ones.

Principal component analysis (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 PC is a linear combination of original variables. The coefficients of each linear combination are obtained from an eigenvector of the 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 the cases 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. Nevertheless, only a limited number of applications of PCA in chemical engineering has been reported in the literature, with hardly any applications to data reconciliation.

In this article, we first give a brief review of the previous work in steady-state linear data reconciliation and gross error detection, followed by a derivation of a set of tests for detect-

Correspondence concerning this article should be addressed to C. M. Crowe.

ing and identifying gross errors based on principal component analysis. We then compare these tests to those already used, and show why the former are sharper than the latter. We also address the problems in the currently used tests. At the end, we present two numerical examples to illustrate how the methods can be applied to practical problems. In particular, we show that the tests are sensitive to subtle gross errors, and have substantially greater power to correctly identify the variables in error than the other tests.

Previous Work in Data Reconciliation and Gross Error Detection

In this section, we briefly review a few basic equations used in steady-state (stationary) data reconciliation, and the currently used statistical tests for detecting gross errors. The residual vector of the process constraints and its covariance matrix that will be defined are the fundamental quantities in all statistical tests.

As shown by Crowe et al. (1983) and Crowe (1986), unmeasured variables in the constraints of a steady-state process, such as mass and energy balance equations, can be removed by the method of matrix projection. Therefore, we can assume, without loss of generality, that unmeasured variables are not present in the constraints. We refer to such constraints as the reduced constraints, and to the constraints involving unmeasured variables as the original constraints. When all variables are measured, the reduced constraints are identical to the original constraints.

Since measurements are contaminated by random errors and frequently by gross errors such as sensor biases and process leaks, the conservation laws and other process constraints are not satisfied. The residuals of reduced constraints for a linear steady-state process can be defined in their matrix form as

$$e = B\tilde{x} \tag{1}$$

where $\tilde{x}$ is a vector of the measured variables. Matrix B is $m \times n$ with $m < n$, known as the balance matrix. It consists of the coefficients of the reduced constraints. Vector e reflects the violation of the constraints by the measurements and is the fundamental vector used in gross error analysis.

If we assume that the measured variables follow a certain distribution, which does not need to be specified at this stage, with the expectation x and the covariance matrix Σ, that is, $\tilde{x} \sim (x, \Sigma)$, such that

$$Bx = 0, \tag{2}$$

then e follows the same distribution with the expectation

$$E(e) \equiv e^* = 0 \tag{3}$$

and the covariance matrix

$$\mathrm{cov}(e) \equiv H_e = B\Sigma B^T \tag{4}$$

in light of Eqs. 1 and 2, that is, $e \sim (e^*, H_e)$.

We know *a priori* that the expectation of e is always zero for measurements with only random errors, whereas that of $\tilde{x}$ is unknown; H_e contains the information of the process structure, expressed in B, and of the measurement covariance structure, expressed in Σ. It is the quantity that jointly captures the process variability and the inherent variability. Together with the vectors e and e^*, it is fundamental in detecting gross errors.

Univariate, maximum power, and chi-square tests

To compare the new tests to those already used in data reconciliation, we briefly review the univariate, MP, and chi-square tests.

Let diag(H_e) be a diagonal matrix whose diagonal elements are as the same as those of H_e. The univariate test for each residual is given by (Reilly and Carpani, 1963)

$$z_{e,i} = \frac{e_i}{\sqrt{(H_e)_{ii}}}, \qquad i = 1, \ldots, m \tag{5}$$

or written collectively,

$$z_e = [\mathrm{diag}(H_e)]^{-1/2} e \tag{6}$$

The MP test is given by (Crowe, 1989)

$$z^*_{e,i} = \frac{(H_e^{-1}e)_i}{\sqrt{(H_e^{-1})_{ii}}}, \qquad i = 1, \ldots, m \tag{7}$$

or written collectively,

$$z^*_e = [\mathrm{diag}(H_e^{-1})]^{-1/2} H_e^{-1} e. \tag{8}$$

The statistics in Eqs. 6 and 8 are special cases of a general test that examines any arbitrary linear combination of the residuals

$$z_{e,i}(V) = \frac{(Ve)_i}{\sqrt{(VH_eV^T)_{ii}}}, \qquad i = 1, \ldots, m \tag{9}$$

or written collectively

$$z_e(V) = [\mathrm{diag}(VH_eV^T)]^{-1/2} Ve. \tag{10}$$

It can be seen that $z_{e,i}$, $z^*_{e,i}$, and $z_{e,i}(V)$ are all unit normal variates if the measurements are normally distributed. The choice of V is arbitrary. The univariate statistics has $V = I$, and the MP statistic has $V = H_e^{-1}$. We will see later that the principal component test also belongs to this class, where V contains the eigenvectors of H_e.

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

$$\chi^2_m = e^T H_e^{-1} e, \tag{11}$$

which examines all residuals together (Reilly and Carpani, 1963).

The univariate, MP, and chi-square tests are frequently used methods in gross error detection. However, there are a few inherent problems associated with them that motivated us to study an alternative method. We address these problems in the section entitled "Relationships Among the Tests."

PC Tests for Residuals of Reduced Process Constraints

Although the information in $\boldsymbol{e}$ can be analyzed in the forms of Eqs. 6, 8, and 11 in detecting gross errors, the analysis can be done differently, in order for us to gain a deeper insight into the problem. In this section, we define a principal component transform of the residual vector of the process constraints, and the tests that examine the principal components both individually and collectively.

Let us consider a set of linear combinations of $\boldsymbol{e}$

$$\boldsymbol{y}_e = \boldsymbol{W}_e^T(\boldsymbol{e} - \boldsymbol{e}^*) = \boldsymbol{W}_e^T \boldsymbol{e}. \tag{12}$$

Vector $\boldsymbol{y}_e$ consists of principal components, and the values of its elements are principal component scores, if the linear combination coefficients given in $\boldsymbol{W}_e$ form eigenvectors of $\boldsymbol{H}_e$, and satisfy

$$\boldsymbol{W}_e = \boldsymbol{U}_e \boldsymbol{\Lambda}_e^{-1/2}. \tag{13}$$

Matrix $\boldsymbol{\Lambda}_e$ is diagonal, consisting of the eigenvalues of $\boldsymbol{H}_e$, $\lambda_{e,i}$, $i = 1, \ldots, m$, on its diagonal, and satisfies

$$\boldsymbol{\Lambda}_e = \boldsymbol{U}_e^T \boldsymbol{H}_e \boldsymbol{U}_e. \tag{14}$$

Matrix $\boldsymbol{U}_e$ consists of the orthonormalized eigenvectors of $\boldsymbol{H}_e$, so that

$$\boldsymbol{U}_e \boldsymbol{U}_e^T = \boldsymbol{I}. \tag{15}$$

It can be shown that $\boldsymbol{y}_e \sim (\boldsymbol{0}, \boldsymbol{I})$ because $\boldsymbol{e} \sim (\boldsymbol{0}, \boldsymbol{H}_e)$. Therefore, a set of correlated variables, $\boldsymbol{e}$, is transformed into a new set of uncorrelated variables, $\boldsymbol{y}_e$. The principal components are numbered in descending order of the magnitudes of the corresponding eigenvalues.

On the other hand, Eqs. 12 and 13 can be combined and rewritten as

$$\boldsymbol{e} = \boldsymbol{e}^* + \boldsymbol{U}_e \boldsymbol{\Lambda}_e^{1/2} \boldsymbol{y}_e \tag{16}$$

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

$$\boldsymbol{e} = \boldsymbol{e}^* + \boldsymbol{U}_e \boldsymbol{\Lambda}_e^{1/2} \boldsymbol{y}_e + (\boldsymbol{e} - \hat{\boldsymbol{e}}) \tag{17}$$

with

$$\hat{\boldsymbol{e}} = \boldsymbol{e}^* + \boldsymbol{U}_e \boldsymbol{\Lambda}_e^{1/2} \boldsymbol{y}_e \tag{18}$$

where $\boldsymbol{y}_e \in \mathrm{R}^{k_e}$, and $k_e < m$. Equation 18 is referred to as the principal component model of $\boldsymbol{e}$. Equation 17 indicates that the residuals in the vector $\boldsymbol{e}$ can be decomposed into the contributions from their expectations, $\boldsymbol{e}^*$ (which are zeros in reconciliation), principal components, $\boldsymbol{y}_e$, and the residuals of the principal component model, $\boldsymbol{e} - \hat{\boldsymbol{e}}$.

If the measured variables are normally distributed, $\tilde{\boldsymbol{x}} \sim \mathrm{N}(\boldsymbol{x}, \boldsymbol{\Sigma})$, we have, from the previous discussion, that

$$\boldsymbol{y}_e \sim N(\boldsymbol{0}, \boldsymbol{I}). \tag{19}$$

Instead of looking at statistical tests for $\boldsymbol{e}$, we can study alternatively how to perform hypothesis testing on $\boldsymbol{y}_e$ and $\boldsymbol{e} - \hat{\boldsymbol{e}}$.

Principal component test

Based on Eqs. 12 and 19, the test for a principal component is defined as

$$y_{e,i} = (\boldsymbol{W}_e^T \boldsymbol{e})_i \sim N(0,1), \qquad i = 1, \ldots, k_e, \tag{20}$$

which can be tested against a threshold tabulated value.

If we substitute V with $\boldsymbol{W}_e^T$ in Eq. 9, and take into account that $\boldsymbol{W}_e \boldsymbol{W}_e^T = \boldsymbol{H}_e^{-1}$, we can see that Eq. 20 is a special case of Eq. 9.

Equation 20 shows that the ith principal component, $y_{e,i}$, is obtained from $\boldsymbol{w}_{e,i}^T \boldsymbol{e}$, where $\boldsymbol{w}_{e,i}$ is the ith eigenvector in $\boldsymbol{W}_e$.

We can identify the constraints in gross error by inspecting the contribution from the jth residual in $\boldsymbol{e}$, e_j, to a suspect principal component, say $y_{e,i}$, which can be calculated by

$$g_j = (\boldsymbol{w}_{e,i})_j e_j, \qquad j = 1, \ldots, m. \tag{21}$$

Define $\boldsymbol{g} = (g_1, \ldots, g_m)^T$, and let $\boldsymbol{g}'$ be the same as $\boldsymbol{g}$ except that its elements are sorted in descending order of their absolute values. We can study the contributions by checking the signs and magnitudes of the elements in $\boldsymbol{g}'$. 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

$$\left| \frac{\left(\sum_{j=1}^{k_1} g_j' \right) - y_{e,i}}{y_{e,i}} \right| \le \epsilon_1 \tag{22}$$

where ϵ_1 is a prescribed tolerance such as 0.1.

It is noted that since the signs of these contributions can be either plus or minus, as can the signs of the elements of $\boldsymbol{w}_{e,i}$ and $\boldsymbol{e}$, the cancellation effect among the elements of $\boldsymbol{g}'$ should be taken into account in identifying the suspect constraints. This is done in Eq. 22.

The calculation 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, and conservative estimates are often used. One such estimate for that of an overall type I error is given by Sidak (1967)

$$\alpha = 1 - (1 - \alpha^*)^m \qquad (23)$$

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

$$\alpha^* = 1 - (1 - \alpha)^{1/m} \approx \frac{\alpha}{m}. \qquad (24)$$

The $\approx$ expression, which leads to the Bonferroni bound (Seber, 1984), holds when $\alpha/m \ll 1$.

When k_e principal components are retained, Eqs. 23 and 24 can be rewritten as

$$\alpha = 1 - (1 - \alpha^*)^{k_e} \qquad (25)$$

$$\alpha^* = 1 - (1 - \alpha)^{1/k_e} \approx \frac{\alpha}{k_e}. \qquad (26)$$

The $\approx$ expression holds when $\alpha/k_e \ll 1$. It is noted that Eq. 25 can be used to calculate the exact probability of the overall type I error from a prescribed α^* for the principal components, and Eq. 26 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.

If m is large, α given by Eq. 23 is also large, while α^* given by Eq. 24 is small. For instance, if $m = 21$, $\alpha^* = 5\%$, we have $\alpha = 65.9\%$. This is simply unacceptable because the chance of a false alarm is about $^2/_3$, even for such a moderate size problem. On the other hand, if $m = 21$, $\alpha = 5\%$, we have $\alpha^* = 0.24\%$, which may lead to an unacceptably high probability of a type II error. This happens particularly when a large process is analyzed.

Since we usually have $k_e < m$, or even $k_e \ll m$ when m is large, it is obvious that α given by Eq. 25 is smaller than its counterpart based directly on e, given by Eq. 23, for any chosen marginal confidence level, $1 - \alpha^*$. Therefore we can in general reduce the overall type I error in detecting gross errors by using the principal component test.

Collective tests

Traditionally, the chi-square collective tests, χ_m^2, which tests the optimal value of the objective function in the model of data reconciliation, is calculated by Eq. 11. Analysis of the causes that inflate this statistic is difficult because of the involvement 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

$$\chi_m^2 = e^T H_e^{-1} e = y_e^T y_e \qquad (27)$$

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

$$\chi_{k_e}^2 = y_e^T y_e, \qquad (28)$$

which can be called the truncated chi-square test, and is a principal component approximation of the chi-square statistic given in Eq. 27.

By employing the principal components, as expressed in Eqs. 27 and 28, we can identify the constraints in gross error by checking the magnitudes of the retained elements of y_e. Since each element in y_e contains information from all the residuals of the process constraints, only the retained principal components need to be considered. Those principal component scores having large absolute values 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'. Equation 22 can still be used to choose the number of the major contributors from the constraints.

It should be noted that the expression of the traditional 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.

Another important collective test statistic is defined by

$$Q_e = (e - \hat{e})^T (e - \hat{e}) \qquad (29)$$

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

$$Q_e = \sum_{i=k_e+1}^{m} \lambda_{e,i} y_{e,i}^2, \qquad (30)$$

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

$$q_1 = \sum_{i=k_e+1}^{m} \lambda_{e,i}, \quad q_2 = \sum_{i=k_e+1}^{m} \lambda_{e,i}^2, \quad q_3 = \sum_{i=k_e+1}^{m} \lambda_{e,i}^3,$$

and

$$h_0 = 1 - \frac{2 q_1 q_3}{3 q_2^2},$$

the upper limit for Q_e can be obtained as

$$Q_{e,\alpha} = q_1 \left(\frac{c_\alpha \sqrt{2 q_2 h_0^2}}{q_1} + \frac{q_2 h_0 (h_0 - 1)}{q_1^2} + 1 \right)^{1/h_0} \qquad (31)$$

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

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

To analyze the causes that inflate the Q_e statistic, similar to the analyses of $\chi^2_{k_e}$, we look at the absolute values of the residuals of the predictions, that is, 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 the elements e_i are highly correlated, as illustrated in the next section.

Let $f = e - \hat{e}$, and f' be as 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

$$1 - \frac{\sum_{i=1}^{k_2} f_i'^2}{Q_e} \leq \epsilon_2 \qquad (32)$$

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. Since each principal component contains the information from all residuals, however, provided that 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 because the principal component model that we used may not fit the changed situation. Therefore, we need to look at the principal components that are not retained in the model, through the Q_e statistic.

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 mainly to take account of the process variability, while leaving the inherent variability of 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, that is, $H_e' = \text{diag}(H_e)$, $\lambda_e \in R^m$ 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

$$k_e = \max(i) | \lambda_{e,i} \geq \lambda'_{e,i}, \qquad i = 1, \ldots, m. \qquad (33)$$

In practice, $k_e \ll m$, 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. 33, because when $\lambda'_{e,i} > \lambda_{e,i}$, the inherent variability becomes dominant, and that is where we should stop adding more principal components.

Relationships among the 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, the MP, and the principal component constraint tests are all identical.

The general test statistic, defined in Eq. 10, reduces to

$$z_e(V) = [VH_eV]^{-1/2}Ve \qquad (34)$$

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

$$z_e(V) = H_e^{-1/2}e \qquad (35)$$

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

For the principal component test statistic defined by Eqs. 12 and 13, we have, when H_e is diagonal, $V = W_e^T = H_e^{-1/2}$, since $U_e = I$ and $\Lambda_e = H_e$. Therefore, V is diagonal, and Eq. 10 again reduces to Eq. 35.

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 hyperrectangular 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 1. This 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. Such a case is of no practical importance, because reconciliation could then be done on the individual components in each unit. The matrix H_e is not diagonal in general, and in

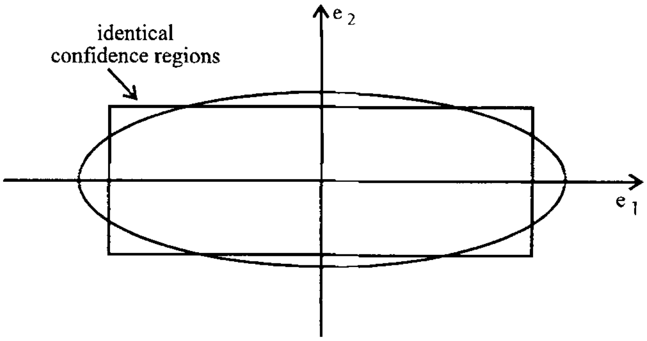

Figure 1. Identical confidence regions of the univariate, MP, and PC tests if H_e were diagonal.

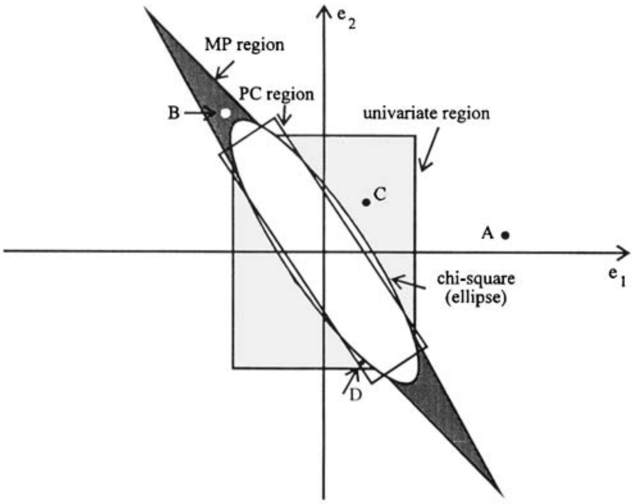

Figure 2. Relationship among z_e, z_e^*, y_e, and χ_m^2 tests when H_e is not diagonal.

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.

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. 6, 8, and 12, in comparison to Eq. 35. An example of the relative positions of the confidence regions of the tests is illustrated in Figure 2 along with the ellipse of the chi-square statistic, when H_e is not diagonal. The inadequacy of using the univariate region under such conditions 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 of the cases among all the tests that we considered in this article, as also illustrated in Figure 2.

We observe from Figure 2 that if there is only one gross error of large magnitude, such as that represented by point A, all tests are able to pick up the unusual event, and the choice of the tests does not matter. For example, when two relatively large gross errors, represented by 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 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 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_m^2 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. 10, including the univariate and the PC tests, but only when there is a single error. When there are more than one gross error, 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 point D. The point is located outside of the MP region, but within the univariate and the PC regions. Figure 2 shows us 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 present below, we show that the principal component tests not only provide better detection even to subtle gross errors, but also have substantial power to correctly identify the variables in error over the other tests.

Practical Issues in Performing Principal Component Tests

Principal component tests provide an insight into a correlated problem that 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 graph and the contribution graphs to identify the variables in gross error. The univariate and MP tests can still be used as additional evidence.

Matrix Σ is crucial in detecting gross errors and in data reconciliation. Before samples are taken all measuring instruments should be carefully checked and calibrated, and the process inspected to have any leaks 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.

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 a set of measurements of a variable would tend toward a normal distribution as the size of the set increased, if they were sampled randomly and independently, because of the central limit theorem. Since measurements are usually correlated in time, however, the central limit theorem may not hold. When the distribution is unknown, enough samples have to be taken to obtain a reference distribution.

There may be some concerns about the principal component tests that we propose. First, the tests work in the probability sense. Regardless of the power of these tests, there is no guarantee that they can always detect gross error if there is at least 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. Second, the principal component tests involve more computation in calculating the eigenvalues and eigenvectors, and more analysis of the contribution graphs.

Test for the Original Constraints and the Adjustments

We have restricted ourselves to the reduced constraints for the sake of simplicity. The relationship between the reduced and the original constraints can be found in Crowe (1992). In principle, if the vector of the residuals of reduced constraints, $\boldsymbol{e}$, is replaced by the vector of measurement adjustments, $\boldsymbol{a}$, or by the vector of the residuals of original constraints, $\boldsymbol{r}$; and $\boldsymbol{H}_e$ is replaced by $\boldsymbol{Q}$ or $\boldsymbol{H}_r$, the covariance matrices of $\boldsymbol{a}$ and $\boldsymbol{r}$, respectively, the test procedures that we presented previously are still valid. However, $\boldsymbol{Q}$ is always singular (Crowe et al., 1983), and $\boldsymbol{H}_r$ is of full rank (identical to $\boldsymbol{H}_e$) only in the absence of unmeasured quantities, but is singular when unmeasured quantities or chemical reactions are present in a process (Crowe, 1986). When $\boldsymbol{Q}$ or $\boldsymbol{H}_r$ is singular, the maximum number of principal components that can be retained, or equivalently, the rank of the corresponding covariance matrix, is less than the number of original constraints or measurements. 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. It can be shown that $\boldsymbol{H}_e$, $\boldsymbol{H}_r$, and $\boldsymbol{Q}$ all have the same rank (Tong, 1995), therefore we restrict

$$k_r \leq \text{rank}(\boldsymbol{H}_r) = \text{rank}(\boldsymbol{H}_e) = m \qquad (36)$$

$$k_a \leq \text{rank}(\boldsymbol{Q}) = \text{rank}(\boldsymbol{H}_e) = m \qquad (37)$$

where k_r and k_a are the numbers of principal components that may be retained in the corresponding tests. In general, k_e, k_r, and k_a may be different.

The test procedures also apply to the nonlinear data reconciliation problem discussed in Crowe (1986). Here, since the matrices, $\boldsymbol{H}_e$, $\boldsymbol{H}_r$, and $\boldsymbol{Q}$, are not constant during reconciliation, the principal component tests should be employed only at the final stage of data reconciliation, where these covariance matrices have their final values.

Numerical Examples

Two simple examples are provided to illustrate how the principal component tests can be used, and compared to the univariate, MP, and chi-square tests. In particular, we will show that the tests are much more sensitive to subtle gross errors, and have substantially greater power to correctly identify the variables in error, than the other tests.

For simplicity, the univariate, MP, and PC tests will be shown on the same graph. Each univariate or MP test corresponds to one of the residuals of the constraints or to one of the measurement adjustments. However, each PC test corresponds to more than one of these, as seen from the corresponding contribution plot. Only the absolute values of the statistics are shown on the graphs.

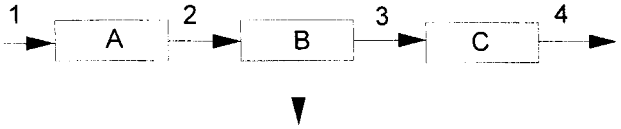

Figure 3. Process with a leak.

A process with a leak

This example was used by Crowe et al. (1983) to demonstrate how chi-square and univariate tests could be used to detect an unsuspected leak. The process consists of three units in series, with an unknown leak in the second unit, as shown in Figure 3. Only total mass balances are considered. In this example the original constraints are identical to the reduced constraints: $\boldsymbol{H}_e = \boldsymbol{H}_r$. The matrices used in the tests can be found in Crowe et al. (1983). In order to make the leak harder to detect, a different set of measurements, $\tilde{x} = (98.4\ 98.6\ 96.5\ 96.2)^T$, was used, giving $\boldsymbol{e} = \boldsymbol{B}\tilde{x} = (-0.2\ 2.1\ 0.3)^T$. The z_e, z_e^*, and y_e statistics are illustrated in Figure 4 at 95% confidence level for each variate, with the threshold value of 1.96, shown on the plot as a dashed line. The conservative estimate of the probability of the overall confidence level for z_e and z_e^* is given by $(1-0.05)^3\% = 85.74\%$. The contributions from each residual of the constraints to the principal components are given in Figure 5. From the figure we see that the residual of the mass balance around unit B is the major contributor to the last principal component, $y_{e,3}$, which is inflated. Since this residual did not contribute to $y_{e,2}$, and $y_{e,2}$ was not an outlier, we can conclude that the imbalance was caused by this residual, and it corresponds to the leak in unit B. The univariate test failed, while the MP and PC tests successfully detected the leak.

If we prefer setting the overall confidence at a higher level, say 95%, the marginal confidence level would be $(1-0.05)^{1/3}\% = 98.3\%$. At such a level, the threshold value for a unit normal variate is 2.39, so that none of the three tests would be violated. The collective statistics are $\chi^2_{m=3} = 4.688(7.82)$, $\chi^2_{k_e=2} = 0.666(5.99)$, and $Q_{e,k_e=2} = \mathbf{2.356}(2.20)$,

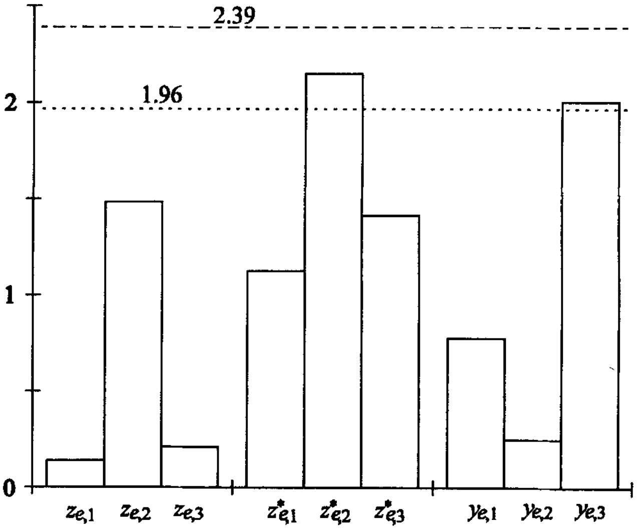

Figure 4. Univariate, MP, and PC tests on the constraints, the process with a leak.

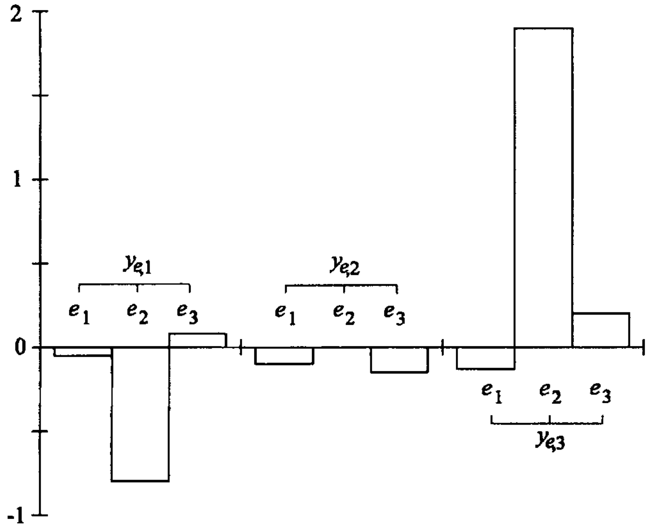

Figure 5. Contributions from the residuals of the constraints to y_e, the process with a leak.

where the numbers in parentheses are the threshold values. We notice that only the Q_e statistic is able to detect the gross error at 95% overall confidence level. This demonstrates the advantage of using the principal component tests to detect a subtle gross error. The contributions from each residual to Q_e are 0.589, **1.178**, and 0.589, respectively, and e_2, which is associated with the leak in unit B, is the major contributor.

Ammonia synthesis loop

The ammonia synthesis system considered by Crowe (1988) is shown in Figure 6. 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 was fixed at 0.02 in our study. Following Crowe, the measurements were generated from the true values by the addition of normally distributed random noise with the same covariance structure as in Crowe et al. (1983). 20% gross error in $NH_3^{(4)}$ and 10% in $N_2^{(1)}$ were added to the corresponding measurements. The true and the measured data were given in Crowe (1988).

First we look at the statistics that are illustrated in Figures 7 and 8. The four reduced constraints correspond to the mole balances of NH_3, N_2, H_2, and Ar, respectively. While z_e in Figure 7 points out that the gross errors are related to the measurements involving NH_3 and H_2, Figure 8 shows that $y_{e,1}$ picked up NH_3 and N_2, $y_{e,2}$ picked up H_2; z_e^* in Figure 7 picked up nothing except H_2. In fact, the measurements of H_2 are not in gross error but are only confounded with that of NH_3. This is an example showing that 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 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, $\chi^2_{k_e=2}$ = **24.94**(5.99), is inflated along with the conventional chi-square test, $\chi^2_{m=4}$ = **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.

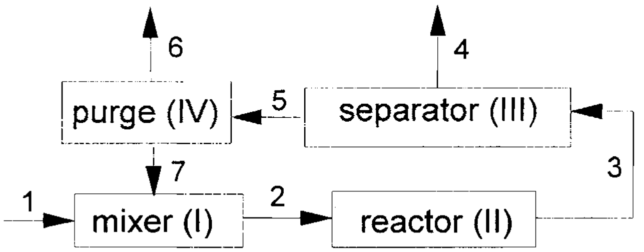

Figure 6. Ammonia synthesis loop.

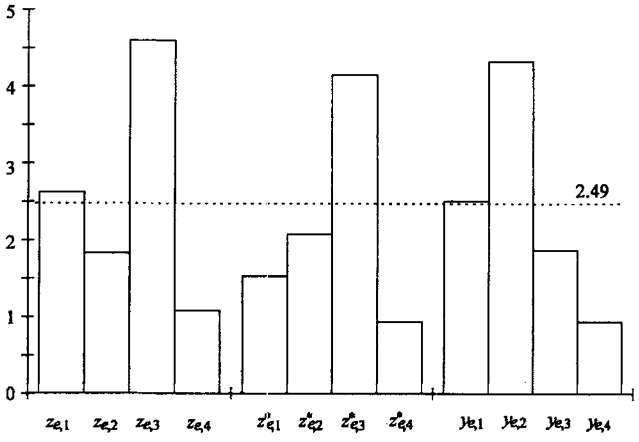

Figure 7. Univariate, MP, and PC tests on the reduced constraints, e: NH_3 loop.

We then look at the tests for $\boldsymbol{r}$. The seventeen original constraints correspond to the component mole balances in all the units of the process, as shown in Table 1, 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.

The z_r, z_r^*, and y_r tests are shown in Figure 9. Some of the univariate statistics are not available because $\boldsymbol{H}_r$ is singular. Although Figure 9 indicates that z_r, z_r^*, and y_r all picked up H_2 and NH_3 (see Table 1 also), the number of the outliers from the first two tests is large, mainly due to confounding among the measured variables caused by the process topology. This complicates any further analysis. The maxi-

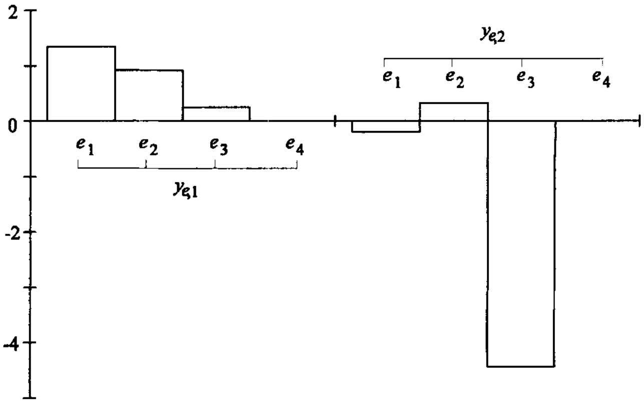

Figure 8. Contributions from the residuals of the reduced constraints to y_e: NH_3 loop.

Table 1. Correspondence Among the Balance Equations and Process Units

Equation No.	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 No.	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 No.	13	14	15	16	17	—
Balance on	$H_2^{(IV)}$	$Ar^{(IV)}$	$N_2^{(s)}$	$H_2^{(s)}$	$Ar^{(s)}$	—

mum number of the principal components that could be retained in this problem is 4. The plot of the contributions to the inflated $y_{r,2}$, shown in Figure 10, 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. 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 plot shown in Figure 11 leads to the same conclusion.

Finally, we look at the tests for $\boldsymbol{a}$: z_a and z_a^* shown in Figure 12 picked up $H_2^{(1)}$, $H_2^{(5)}$, and $NH_3^{(4)}$. Figure 13 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 plot shown in Figure 13 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.

Once we know the measurements are contaminated by gross errors, we have to identify the measurements that are in error, and remove them before 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_m^2) 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 article.

We shall see that the principal component 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 $\{N_2^{(1)}, NH_3^{(4)}\}$ or the secondary suspect $\{H_2^{(1)}, N_2^{(3)}\}$ passed the univariate and the MP tests for the reduced constraints and for 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 2 summarizes the results of all the collective tests (chi-square, truncated chi-square, and Q tests) on $\boldsymbol{e}$ and $\boldsymbol{a}$ when

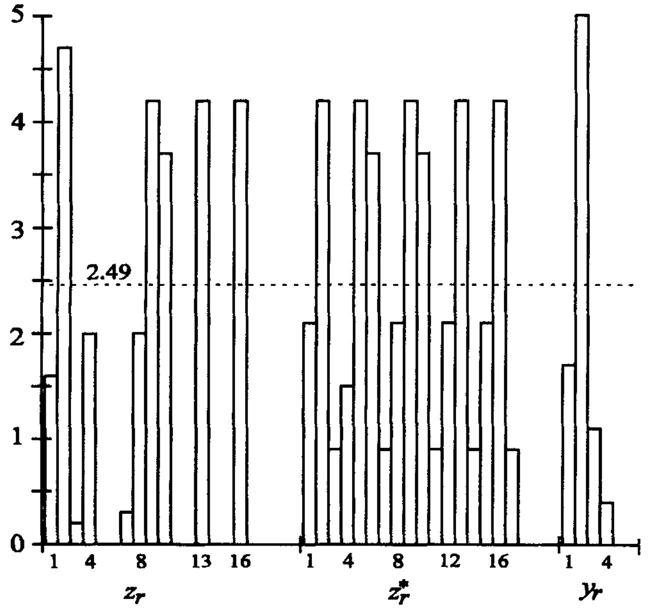

Figure 9. Univariate, MP, and PC tests on the original constraints, r: NH_3 loop.

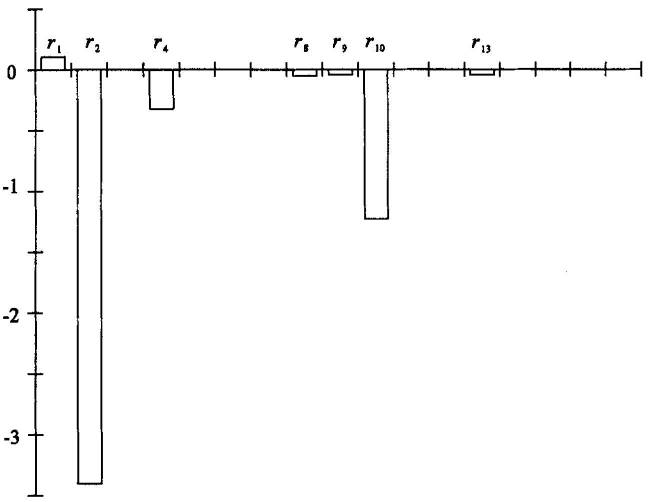

Figure 10. Contributions from the residuals of the original constraints to $y_{r,2}$: NH_3 loop.

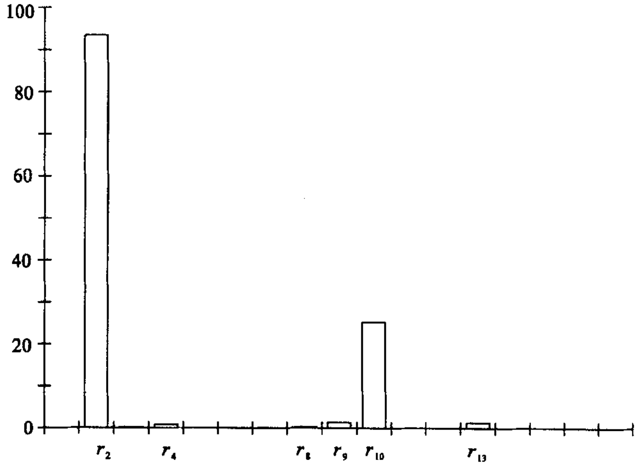

Figure 11. Contributions from the residuals of the original constraints to Q_r: NH_3 loop.

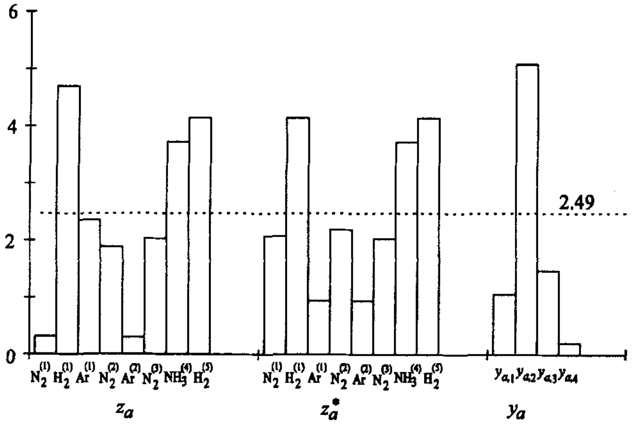

Figure 12. Univariate, MP, and PC tests on the measurement adjustments, a: NH_3 loop.

Table 2. Collective Tests for the Suspect Pairs*

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

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

Table 3. Inflated Test Statistics with $NH_3^{(4)}$ Deleted ($k_r = k_a = 2$)

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

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 principal component collective statistics are able to distinguish the three pairs, and correctly identify the pair $\{N_2^{(1)}, NH_3^{(4)}\}$, which is really in gross error.

To further demonstrate the advantage of using principal component 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 14. 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 to gross error, while $H_2^{(1)}$ was picked up because of its confounding with $NH_3^{(4)}$. Without using the principal component test, the identify 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 3. We noted that no univariate and maximum power test statistics were significant.

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 in any test statistics. 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.

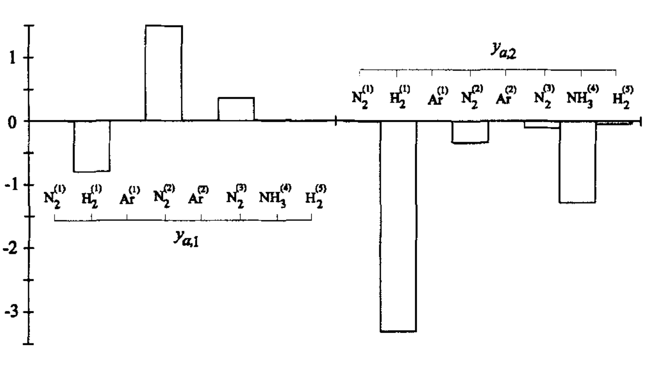

Figure 13. Contributions from the measurement adjustments to $y_{a,1}$ and $y_{a,2}$: NH_3 loop.

Conclusions

A new set of test statistics, principal component tests, has been derived based on principal component analysis, and compared to the univariate, MP, and chi-square tests. The PC statistic is identical to the univariate and MP statistics only under an unrealistic limiting condition, but in general it is sharper. The analysis of the contributions to the principal components from individual variables is useful in identifying the variables in gross error. The use of the PC statistics is helpful in controlling the type I error because the correlation among variables is removed. Two collective statistics, χ^2_{ke} and Q_e, are derived from principal component analysis. The former examines the retained principal components, and the latter examines the unretained ones. Both of them can be traced down to the individual contributions from the residuals of the process constraints, which is helpful in isolating and identifying gross errors, an advantage over the conventional chi-square statistic. Similar statistics were derived for the original constraints and the measurement adjustments. The number of principal components retained in the principal component

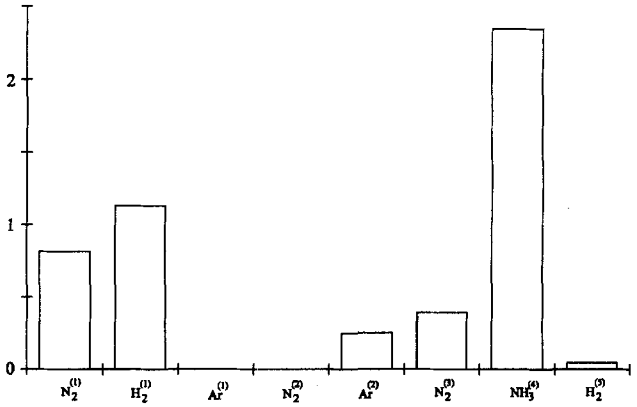

Figure 14. Contributions from measurement adjustments to Q_a: NH_3 loop (7.1% gross error level).

model is often far less than that of the independent original variables in a large problem, hence the dimension at the monitoring and detecting level may be greatly reduced. Our study showed that the principal component test is very useful in detecting subtle gross errors when the other tests fail. We also illustrated that as correlation increases, the univariate and the MP tests are not reliable, while the PC test provides a consistent performance. The MP test should be used with caution when correlation is high due to its flattened parallelogram confidence region.

Acknowledgment

The authors are grateful to the Natural Sciences and Engineering Research Council of Canada for a Research Grant in support of this work.

Notation

I = identity matrix
m = rank of $\boldsymbol{H}_e$ and the number of reduced process constraints
n = number of measured variables
$w_{e,i}$ = ith eigenvector of $\boldsymbol{H}_e$
λ_e, λ_e' = vectors of eigenvalues of $\boldsymbol{H}_e$ and $\boldsymbol{H}_e'$, respectively

Literature Cited

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

Crowe, C. M., "Data Reconciliation—Progress and Challenges," *Proc. Int. Symp. on Process Sys. Eng.*, **1**, 111, Kyongju, Korea (May 30–June 3, 1994).

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

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

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).

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).

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

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

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

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

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

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

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

Nomikos, P., and J. F. MacGregor, "Monitoring Batch Processes Using Multiway Principal Component Analysis," *AIChE J.*, **40**, 1361 (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. Macquin, and D. Sauter, "Data Validation Using Orthogonal Filters," *IEE Proc.-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, Que. (Oct., 1963).

Seber, G. A. F., *Multivariate Observations*, Wiley, New York (1984).

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

Tamhane, A. C., "A Note on the Use of Residuals for Detecting an Outlier in Linear Regression," *Biometrika*, **69**, 488 (1982).

Tong, H., "Studies in Data Reconciliation and the Detection of Gross Errors," PhD Thesis, in preparation, McMaster University, Hamilton, Ont., Canada (1995).

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

Wold, S., K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics Intell. Lab. Syst.*, **2**, 37 (1987).

Manuscript received Jan. 19, 1994, and revision received Sept. 2, 1994.