Chem. Eng. Comm., 2000, Vol. 183, pp. 119–139
Reprints available directly from the publisher
Photocopying permitted by license only

PERFORMANCE EVALUATION OF PCA TESTS IN SERIAL ELIMINATION STRATEGIES FOR GROSS ERROR IDENTIFICATION

MIGUEL BAGAJEWICZ[a,*], QIYOU JIANG[a] and MABEL SÁNCHEZ[b]

[a]*School of Chemical Engineering and Materials Science, University of Oklahoma, Norman, OK, USA;* [b]*Planta Piloto de Ingeniería Química (UNS – CONICET), Camino La Carrindanga. Km 7, (8000) Bahía Blanca–Argentina*

(Received 28 December 1999; In final form 10 April 2000)

In this paper, the performance of the Principal Component Measurement Test (PCMT) is evaluated when used for the identification of multiple biases. A serial elimination strategy is implemented where a statistical test based on principal component analysis is used to identify the measurement to eliminate. A simulation procedure involving random measurement errors and fixed gross error sizes is applied to evaluate its performance. This performance is compared with the one obtained using serial elimination using the conventional Measurement Test (MT), as it is performed in some commercial simulators. The analysis indicates that principal component tests alone, without the aid of other collective tests, do not significantly enhance the ability in identification features of this strategy, performing worse in some cases. A few cases of severe failure of this strategy are shown and a suggestion to test other strategies is offered.

Keywords: Data reconciliation; Gross errors; Principal component analysis

INTRODUCTION

Data reconciliation techniques are performed to estimate process variables in such a way that balance constraints are satisfied, but to obtain accurate estimates, some action should be taken to eliminate the influence of gross errors. Hypothesis testing has been extensively used for detection and

*Corresponding author. Tel.: 4053255458, Fax: 4053255813, e-mail: bagajewicz@ou.edu

identification purposes. A good survey of this issue can be found in Mah (1990) and Madron (1992).

For the purpose of identification of gross errors, Mah (1982) proposed the measurement test. This test has been used in gross error hypothesis testing in almost all practical applications. Narasimhan and Mah (1987) introduced the GLR test, which can also provide an estimate of the gross error. Crowe (1988) proved that these two are equivalent and have maximum power for the case of one gross error. Recently Principal Component Tests (PCT) have been proposed by Tong and Crowe (1995, 1996) as an alternative for multiple bias and leak identification. Industrial applications of PCMT were reported by Tong and Bluck (1998), who indicated that tests based on PCA are more sensitive to subtle gross errors than others, and have larger power to correctly identify the variables in error than the conventional Nodal, Measurement and Global Tests. With the exception of the aforementioned scattered examples of improved efficiency for one gross error, there is no published work assessing the effectiveness of the method for multiple gross errors.

For the case of multiple gross errors, serial elimination was originally proposed by Ripps (1965). The method proposes to eliminate the measurement that renders the largest reduction in a test statistics until no test fails. Several authors proposed different variants of this procedure (Nogita, 1972; Romagnoli and Stephanopoulos, 1981; Crowe *et al.*, 1983; Rosenberg *et al.*, 1987). Of all these strategies, the one that has been implemented in commercial software more often is the one where the largest measurement test (MT) is used to determine the measurement to eliminate.

The increasing application of PCA in process monitoring and fault diagnosis, and the lack of performance evaluation studies like those proposed by Iordache *et al.* (1985), Serth and Heenan (1986), motivates the present work. It involves a comparison of the identification performance of PCMT and MT for existing industrial plants, where multiple gross errors are present. The strategy presented does not include recommendations performed by Tong and Crowe (1995) to complement the usage of the PCMT with information from collective tests. Incorporating such tests in our scheme would not constitute a fair comparison, neither for the usage of PC, nor for the usage of the measurement test.

The paper is organized as follows. First, the PCT is reviewed followed by a short description of the serial elimination strategy and the new strategy tested. Finally, performance evaluation results and a comprehensive discussion are provided.

PRINCIPAL COMPONENT TESTS

In this section, a review of principal component tests proposed by Tong and Crowe (1995) is presented. The Principal Component Nodal Test (PCNT) and the Principal Component Measurement Test (PCMT) are briefly described.

Given a linear steady-state process, the residuals of the constraints are defined as

$$r = Ay \tag{1}$$

where y is the vector of measurements and A the balance matrix. For simplicity, in this formulation it is assumed that all variables are measured. This is not a limitation since the unmeasured variables can be removed using, for example, the Reduced Balance Scheme (Vaclavek, 1969), Matrix Projection (Crowe, 1983), QR decomposition (Schwartz, 1989); (Sánchez and Romagnoli, 1996) or Matrix co-optation (Madron, 1992).

Assuming that the measurement errors follow a certain distribution with covariance Ψ then r will follow the same distribution with expectation and covariance given by

$$E\{r\} = r^* = 0 \quad \Phi = \text{cov}\{r\} = A\Psi A^T \tag{2}$$

The eigenvalue decomposition of matrix Φ is formulated as

$$\Lambda_r = U_r^T \Phi U_r \tag{3}$$

where,

U_r: matrix of orthonormalized eigenvectors of $\Phi (U_r U_r^T = I)$
Λ_r: diagonal matrix of the eigenvalues of Φ.

Accordingly, the following linear combinations of r are proposed

$$p^r = W_r^T (r - r^*) = W_r^T r \tag{4}$$

where,

$$W_r = U_r \Lambda_r^{-1/2} \tag{5}$$

and p^r consists of the principal components. Also, $r \sim P(0, \Phi) \Rightarrow p^r \sim P(0, I)$, for any distribution P. That is, a set of correlated variables r is transformed into a new set of uncorrelated variables p^r. If the measurement errors are normally distributed then the principal components will be normally distributed too. That is: $y \sim N(x, \Psi) \Rightarrow p^r \sim N(0, I)$. Consequently, instead of

looking at a statistical test for r, the hypothesis test may be performed on p^r. Tong and Crowe (1995) proposed the following Principal Component Nodal Test:

$$p_i^r = (W_r^T r)_i \sim N(0,1) \quad i = 1, \ldots, npr \tag{6}$$

which is tested against a threshold tabulated value. The constraints suspected to be in gross error can further be identified by looking at the contribution from the jth residual in r (r_j) to a suspect principal component, say p_i^r, which can be calculated by

$$g_j = (w_i^r)_j r_j \quad j = 1, \ldots, m \tag{7}$$

where w_i^r is the ith eigenvector in W_r.

Similarly, the Principal Component Measurement Test can be stated as follows: The vector of adjustments a and its covariance matrix V are

$$a = \Psi A^T (A \Psi A^T)^{-1} A y \tag{8}$$

$$V = \Psi A^T (A \Psi A^T)^{-1} A \Psi \tag{9}$$

The vector of principal components, p^a, is the following linear combination of a

$$p^a = W_a^T (a - a^*) = W_a^T a \tag{10}$$

Since V is singular, Λ_a has some eigenvalues that are zero. Then, instead of normalizing the whole vector p^a and apply the test, as in the case of the PCNT, we only normalize those components whose eigenvalues are different from zero. Without loss of generality, consider that the first t eigenvalues are different from zero. Then define

$$w_{a,i} = U_{a,i} \lambda_{a,i}^{(-1/2)} \quad i = 1, \ldots, t \tag{11}$$

where $\lambda_{a,i}$ is the eigenvalue, that is, the ith diagonal position of

$$\Lambda_a = U_a^T V U_a \tag{12}$$

Thus, the normalized principal components are:

$$p_i^a = (w_i^a)^T a \quad i = 1, \ldots, t \tag{13}$$

Then, $a \sim P(0, V) \Rightarrow p^a \sim P(0, I)$. When the adjustments are normally distributed, Tong and Crowe (1995) proposed the Principal Component Measurement Test (PCMT) to be based on the testing of the t uncorrelated variables p_i^a against a threshold tabulated value.

In both PCNT and PCMT, the measurements in gross errors can further be identified by looking at the contribution from the jth residual/adjustment to a suspect principal component, say, p_i^a. This contribution is calculated as follows:

$$g_j = (w_i^a)_j a_j \quad j = 1, \ldots, n \tag{14}$$

To assess the number of major contributors k_1 for a suspect principal component p_i^a, a vector g' is defined that contains the elements g_j in descending order of their absolute values. Then k_1 is set so that

$$\left| \frac{\sum^{k_1} g_j' - p_i^a}{p_i^a} \right| \leq \xi \tag{15}$$

where ξ may be fixed, for example at 0.1.

REVIEW OF SERIAL ELIMINATION STRATEGY

Serial elimination was originally proposed by Ripps (1965). The method proposed to eliminate the measurement that renders the largest reduction in the Chi-squared statistics and then repeat the procedure eliminating measurements one at a time until the maximum number of measurements allowed to be deleted is reached or the statistics falls below the threshold value. Nogita (1972) later modified this approach by proposing to eliminate the measurement that reduces the objective function the most and stopping when the objective function increases or the maximum of deletions has been reached. Romagnoli and Stephanopoulos (1981) proposed a method to re-evaluate the objective function without having to solve the reconciliation again and obtain an estimate of the gross errors. Additionally, Crowe *et al.* (1983) proposed a strategy where the global test is used in conjunction with other test statistics. Rosenberg *et al.* (1987) presented an extension of the method where instead of only one measurement being deleted at a time, sets of measurements of different size are deleted. We now present Rosenberg *et al.*'s version of the algorithm followed by the version of Serial Elimination based on the measurement test that is used in many commercial packages.

Serial Elimination Strategy Based on the Global Test (SEG)

The basis of the method is that measurements are deleted sequentially in groups of size d, $t = 1, 2 \ldots, d_{\max}$. After each deletion, the process constraints

are projected (matrix A is recalculated) and the global test is again applied. This procedure continues until among all sets of measurements that are deleted one can find one for which the global test indicates no gross error, or until d_{max} is reached. In this last case, the set that produces the largest reduction in test statistics is declared as the set of gross errors.

Let the following be defined:

S: set of all measured stream flow rates

S_d: temporary set which contains the measurements being deleted in a particular step,

S_c: current set of measurements suspected of containing gross errors.

d_{max}: maximum number of measurements that can be deleted from the original network so that the degree of freedom of the constraints is equal to one (rank $A = 1$). This is equivalent to leave the system with one degree of redundancy.

F_{min}: area under the Chi-square distribution that represents the desired degree of confidence.

The gross error identification steps are:

A. Determine d_{max}. Set $d=0$ and $S_c=0$.

B. Determine the degrees of freedom, ν, for the network containing the measurement set S, and calculate the global test statistic, $\tau = r^T \Phi^{-1} r$. If $\tau < \chi^2_{\nu,\alpha}$, declare no gross errors and stop. Otherwise, go to Step C.

C. Set $d=d+1$. If $d > d_{max}$ declare all measurements in S suspect and stop. Otherwise, set $F_{min}=1-\alpha$. For each possible combination S_d of d measurements, do the following:

   (a) Delete the set S_d from the network.

   (b) Obtain the new constraint matrix A to reflect the reduced set of measurements, $S-S_d$. If one or more columns of the new matrix A are zero, discard this set S_d and choose the next set of d measurements.

   (c) Determine the degrees of freedom ν for the projected constraints, *i.e.*, the rank of A. Calculate the global test statistic τ for the projected constraints.

   (d) If $F\{\chi^2_\nu\} < F_{min}$, replace S_c by S_d and reset $F_{min} = F\{\chi^2_\nu\}$. If all sets of d measurements have been tried, go to D. Otherwise, choose the next set of d measurements in Step (a).

D. If $F_{min} < 1-\alpha$, declare all measurements in S_c in suspect and stop. Otherwise, go to Step C.

Serial Elimination Strategy Based on the Measurement Test (SEM)

The serial elimination strategy based on measurement test (SEM) is obtained based on the modification of SEG. The gross error identification steps are:

A. Determine d_{max}. Set $d=0$ and $S_c=0$.
B. Run the data reconciliation and calculate the measurement test (MT) statistics. If no MT flags, declare no gross errors and stop. Otherwise, go to Step C.
C. If the number of elements in $S_c > d_{max}$ declare all measurements in S_c suspect and stop. Otherwise, put the stream with the largest MT into S_c and do the following:
   (a) Delete the set S_c from the network.
   (b) Obtain the new constraint matrix A to reflect the reduced set of measurements, $S-S_c$.
   (c) Do the data reconciliation and calculate MT for the new redundant system.
D. If no MT flags, declare all measurements in S_c in suspect and stop. Otherwise, go to Step C.

MODIFICATION OF SEM TO USE PCMT

In order to compare the performance of MT and PCMT in the serial elimination strategy, the serial elimination strategy based on PCMT, which is called PCMT–SEM in this paper, is proposed.

The procedure is:

A. Determine d_{max}. Set $d=0$ and $S_c=0$.
B. Run the data reconciliation and calculate the PCMT statistics. If no PCMT flags, declare no gross errors and stop. Otherwise, go to Step C.
C. If the number of elements in $S_c > d_{max}$, declare all measurements in S_c suspect and stop. Otherwise put the stream with the largest contribution to the largest PCMT into S_c and do the following:
   (a) Delete the set S_c from the network.
   (b) Obtain the new constraint matrix B to reflect the reduced set of measurements, $S-S_c$.
   (c) Do the data reconciliation and calculate PCMT for the new redundant system.
D. If no PCMT flags, declare all measurements in S_c in suspect and stop. Otherwise, go to Step C.

SIMULATION PROCEDURE AND UNCERTAINTY REMOVAL

A simulation procedure was applied to evaluate the performance of the aforementioned strategies. The method proposed by Iordache *et al.* (1985) was followed. Each result is based on 10000 simulation trials where the random errors are changed and the magnitudes of gross errors are fixed.

Three performance measures are used: overall power (OP), average number of Type I errors (AVTI) and expected fraction of perfect identification (OPF). They are defined as follows:

$$\text{OP} = \frac{\text{No. of gross errors correctly identified}}{\text{No. of gross errors simulated}} \tag{16}$$

$$\text{AVTI} = \frac{\text{No. of gross errors incorrectly identified}}{\text{No. of simulation trials}} \tag{17}$$

$$\text{OPF} = \frac{\text{No. of trials with perfect identification}}{\text{No. of simulation trials}} \tag{18}$$

The first two measures are proposed by Mah and Narasimhan (1987) and the last one by Rollins and Davis (1992).

A set of gross errors may have its equivalent sets, as described by Bagajewicz and Jiang (1998). Thus, to assess these uncertainties, a new measure, the overall performance of equivalent identification (OPFE) was introduced recently (Sanchez *et al.*, 1999).

$$\text{OPFE} = \frac{\text{No. of trials with successful identification}}{\text{No. of simulation trials}} \tag{19}$$

Determination of OPFE

The uncertainty in gross error detection was discussed in a recent paper (Bagajewicz and Jiang, 1998). Two sets of gross errors are considered equivalent when they have the same effect in data reconciliation. Equivalent sets usually have the same gross error cardinality. However, in some cases when a set of gross errors has special sizes (usually equal to each other) it can be represented by another set of gross errors with different cardinality. These cases are called Degenerate.

When a set of gross errors is obtained, one can identify if it is a successful identification by simply applying the conversion equation between equivalent

sets, which has been proposed by Jiang and Bagajewicz (1999):

$$[AL_1\ K_1]\begin{bmatrix}\hat{\delta}_1\\ \hat{\gamma}_1\end{bmatrix} = [AL_2\ K_2]\begin{bmatrix}\hat{\delta}_2\\ \hat{\gamma}_2\end{bmatrix} \tag{20}$$

where A is the incidence matrix, $\hat{\delta}_1, \hat{\gamma}_1$ vectors of biases and leaks for the set of gross errors identified, $\hat{\delta}_2, \hat{\gamma}_2$ vectors of biases and leaks for the set of gross errors introduced, L_1, K_1, L_2, K_2 matrices reflecting the positions of biases and leaks in the system.

Pre-multiplying both $[AL_1\ K_1]$ and $[AL_2\ K_2]$ by a certain particular matrix, one can transform $[AL_2\ K_2]$ into a canonical form, and obtain the new gross error sizes $\hat{\delta}_2$ and $\hat{\gamma}_2$.

In addition, sometimes many sets of gross errors can represent degeneracy if certain tolerance is allowed. These situations are called Quasi-Degeneracy (Jiang and Bagajewicz, 1999). For example, consider the flowsheet of Figure 1. In particular consider one existing gross error in S_2 of size $\delta_2 = -1$. Consider now that a particular gross error identification method finds gross errors in S_4 and S_5 of sizes $\delta_4 = +1$, $\delta_5 = +1$. These variables are part of the equivalent set (S_2, S_4, S_5), which has gross error cardinality 2. To determine whether the identification is successful one should be able to convert from the set of gross errors found to the originally introduced. In this case, this is possible because, by virtue of degeneracy, the two identified gross errors are equivalent to $\delta_2 = -1$.

Quasi-degeneracy takes place when for example the gross errors found are of sizes $\delta_4 = +0.98$, $\delta_5 = +1.01$. Strictly, this set does not represent a degenerate case. Rather, the conversion to an equivalent set containing S_2, as for example (S_2 S_5), gives values $\delta_2 = -0.98$, $\delta_5 = -0.01$. Therefore, if δ_5 is ignored because its size is too small compared to a tolerance, the gross error introduced are retrieved and it can be claimed that the identification was successful.

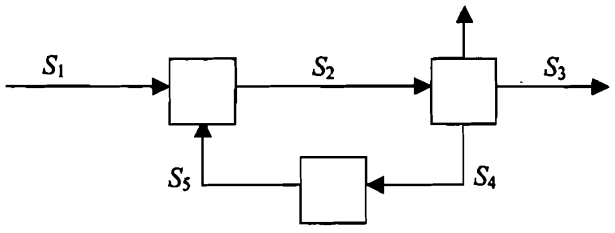

FIGURE 1

In Eq. (20), both sides are vectors. Strictly, when quasi degeneracy is not allowed, both sides have to be equal to declare a successful identification. When Quasi-degeneracy is allowed, both sides are compared within certain threshold tolerance ε_D.

Quasi Equivalency

Consider the following example. Assume that in Figure 1, a gross error is introduced in stream S_1 of size $\delta_1 = +1$. Assume also that the gross error identification finds two gross errors in S_1 and S_2, with sizes $\delta_1 = +0.98$, $\delta_2 = +0.05$. This is a type I error, but accompanied with a small size estimate. In principle, although the result is based on the usage of statistical tests, one is tempted to disregard δ_2 and declare the identification successful. One important observation in this case is that S_1 and S_2 are not a basic set of any subset of the graph. In other words, no degeneracy or equivalency can apply.

Thus, generalizing Quasi-equivalency occurs when only a subset of the identified gross errors is equivalent to the introduced gross errors, and in addition, the nonequivalent gross errors are of small size. Quasi-equivalency is also detected using Eq. (20) and a threshold tolerance ε_E.

OPFE is calculated in this paper by allowing both Quasi-Degeneracy and Quasi-equivalency.

RESULTS

The process flowsheet in Figure 2 is used with comparative purposes. It consists of a recycle system with five units and nine streams. The true flow rate values are $x = [10.\ 20.\ 30.\ 20.\ 10.\ 10.\ 10.\ 4.\ 6.]$. The flow rate standard deviations were taken as 2% of the true flow rates.

Measurement values for each simulation trial were taken as the average of ten random generated values. In order to compare results on the same basis,

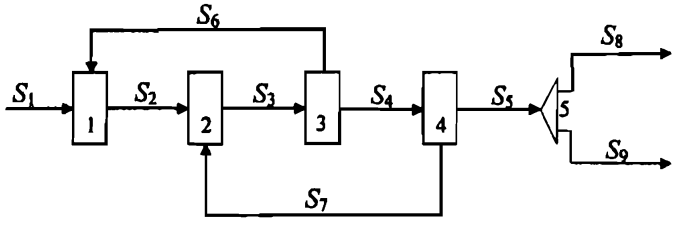

FIGURE 2 An example.

the level of significance of each method was chosen such that it gives an AVTI equal to 0.1 under the null hypothesis when only random errors are present. This practice was first proposed by Mah and coworkers (Rosenberg *et al.*, 1987) so that mispredictions do not count against a method. Nevertheless, the comparison was also repeated using a common confidence level for all methods (95%) revealing slight changes that do not alter the conclusions.

Each strategy was tested under the same scenarios of gross errors introduced. The size of gross error is selected as four times of its corresponding flowrate standard deviation when there is only one gross error, five and three times for two gross errors, five, four and three times for three gross errors and five, four, three and two times for four gross errors. If a leak is introduced, then the minimum standard deviation of flowrates connected to the corresponding process is chosen for the selection of the leak's size. In many papers where the power of gross error detection is assessed, it has been customary to use large sizes (around 10 standard deviations). In this paper, we have chosen relatively smaller values, especially because of the claim that PCA is more efficient in the case of small gross errors. The thresholds ε_D and ε_E were chosen as 10% of the minimum gross error size in the introduced set.

Table I is a comparison between the two different tests MT and PCMT before any gross error identification strategy is incorporated. It indicates that in 6 out of 14 cases, PCMT gets much lower OP than SEM and in the rest cases, they are almost the same.

TABLE I Performance comparison between MT and PCMT

Gross error introduced			*SEM*		*PCMT*	
No.	*Location*	*Size*	*AVTI*	*OP*	*AVTI*	*OP*
	S1	0.8	3.2795	1.0000	2.4760	1.000
	S3	2.4	2.6343	1.0000	0.1955	1.0000
1	S5	0.8	3.2787	1.0000	2.4736	1.0000
	S7	0.8	2.1114	0.9999	2.6201	0.4217
	S1,S4	1.0,1.2	5.0939	1.0000	2.8565	1.0000
	S3,S8	3.0,0.24	5.0432	0.8035	3.0168	0.5000
2	S6,S7	1.0,0.6	4.0795	1.0000	3.0501	0.4974
	S2,S5	2.0,0.6	5.1454	1.0000	3.1049	1.0000
	S1,S3,S5	1.0,2.4,0.6	5.0366	0.9553	2.2265	0.9987
	S2,S4,S8	2.0,1.6,0.24	4.9657	0.6757	2.1044	0.6667
3	S6,S7,S9	1.0,0.8,0.36	4.1343	1.0000	4.8341	0.3974
	S1,S2,S4,S5	1.0,1.6,1.2,0.4	4.5580	0.7597	1.9962	0.8514
	S4,S6,S7,S8	2.0,0.8,0.6,0.16	2.7212	0.8590	3.3766	0.2506
4	S8,S5,S3,S1	0.4,0.8,1.8,0.4	2.3762	0.6249	1.8055	0.6966

Note: S_n means a bias in stream S_n.

Table II shows the comparison between SEM and PCMT–SEM. When there is one gross error, both of them reached high and similar performance. For more than one gross errors, the OPF for the two methods drops to low values, reaching zero for the most cases of four gross errors. However, the OPFE for PCMT–SEM is slightly higher in most cases. The AVTI of all these runs is substantially large for the case of many gross errors. This indicates an average large Type I error. However, these are based on considering the criterion of perfect identification. Within these "failures" there are gross errors that are equivalent to the introduced. A different version of AVTI in which all equivalent and degenerate identifications are considered as successful would show lower values consistent with the large OPFE scores.

The overall conclusion that one can make after examining all these performance measures is that PCA is of help in some situations and it does not make a difference in others. Moreover, it performs poorly in some cases.

The second example is a Large Plant example. It consists of 93 streams (3 of them are unmeasured: S46, S49 and S50), 11 processes, 14 tanks and 9 nodes. Real data was used to perform data reconciliation and then the reconciled data were used as true values in our experiments. This adjusted data are shown in the Appendix. The measurement values for each simulation trial were taken as the average of 10 random generated values. Considering the time expense for this large system, in this experiment each result is based on 1000 (rather than 10000) simulation trials where the random errors are changed and the magnitudes of gross errors are fixed.

Table III indicates that in 5 out of the total 7 cases PCMT–SEM was successful and got a slightly higher performance than SEM. However, there are two cases that PCMT–SEM completely failed, while SEM was still successful.

The failure can be explained inn light of the PCA strategy. The assumption that the variable with larger contribution to the larger principal component has larger probability of having a gross error is not always true. We now show the details of case 4, which illustrate this assertion.

(1) *First iteration*

(a) Number of principal components: 31

(b) Suspect principal components and their corresponding major contributions (both principal components and contributions are sorted from large to small absolute size):

Principal Component	Major Contributions
23	S15, S14
8	S74, S24, S23, S88
27	S13, S14, S12, S15
25	S14, S15, S26
20	S14, S15, S36, S29

(c) List of candidates: S15, S14, S74, S24, S23, S88, S13, S12, S26, S36, S29

(d) Stream to be deleted: S15.

(2) *Second iteration*

(a) Number of principal components: 30

(b) Suspect principal components and their corresponding major contributions:

Principal Component	Major Contributions
23	S14
8	S74, S23, S24, S88
28	S13, S14, S12
24	S14
20	S14, S29, S36, S71

(c) List of candidates: S14, 574, S23, S24, S88, S13, S12, S29, S36, S71

(d) Stream to be deleted: S14.

(3) *Third iteration*

(a) Number of principal components: 29

(b) Suspect principal components and their corresponding major contributions:

Principal Component	Major Contributions
8	S74, S23, S24, S88

(c) List of candidates: S74, S23, S24, S88

(d) Stream to be deleted: S74.

In this step, the size of principal component 8 is -7.077 and the contributions from S74 and S23 are -2.793 and -1.377 respectively. The wrong candidate S74 is selected.

(4) *Fourth iteration*

(a) Number of principal components: 28
(b) Suspect principal components and their corresponding major contributions:

Principal Component	Major Contributions
8	S24, S23

(c) List of candidates: S24, S23
(d) Stream to be deleted: S24.

(5) *Fifth iteration*

(a) Number of principal components: 27
(b) Suspect principal components and their corresponding major contributions:

Principal Component	Major Contributions
13	S37, S29, S17, S34, S79, S30, S42

(c) List of candidates: S37, S29, S17, S34, S79, S30, S42
(d) Stream to be deleted: S37.

(6) *Sixth iteration*

(a) Number of principal components: 26
(b) Suspect principal components: None.

(7) Identified gross errors: S15, S14, S74, S24, S37.

Remark 1 The failure of case 4 *also indicates a case of failure of identification of one single component*. Indeed, if the measurements in streams S14 and S15 are eliminated and the gross error in S23 is left, then the system has only one gross error. The principal component test will have S23 in it suspect candidate, but a serial elimination strategy would fail to identify it as the variable in gross error.

Remark 2 In the absence of other tests, not picking the largest contribution to the largest principal component has a combinatorial method as an alternative. If all the contributions need to be considered, a procedure has to be implemented to sort out this list of candidates and identify one.

TABLE II Performance comparison between SEM and PCMT–SEM

Gross error introduced			*SEM*				*PCMT–SEM*			
No.	*Location*	*Size*	*AVTI*	*OP*	*OPF*	*OPFE*	*AVTI*	*OP*	*OPF*	*OPFE*
	S1	0.8	0.0910	1.0000	0.9107	0.9107	0.0881	1.0000	0.9167	0.9167
	S3	2.4	0.0793	1.0000	0.9225	0.9225	0.0878	1.0000	0.9187	0.9187
1	S5	0.8	0.0939	1.0000	0.9077	0.9077	0.0917	1.0000	0.9122	0.9122
	S7	0.8	0.0984	0.9786	0.9057	0.9065	2.0267	0.0000	0.0000	0.8594
	L3	0.8	2.0658	0.0000	–	0.9157	3.0131	0.0000	–	0.8712
	S1,S4	1.0,1.2	0.0725	0.9988	0.9300	0.9313	0.0745	1.0000	0.9296	0.9296
	S3,S8	3.0,0.24	0.1045	0.9526	0.8500	0.8500	2.2411	0.5000	0.0000	0.1190
2	S6,S7	1.0,0.6	0.1580	0.9444	0.8437	0.8461	3.0680	0.0000	0.0000	0.8929
	S2,S5	2.0,0.6	0.0788	1.0000	0.9219	0.9219	0.0811	1.0000	0.9225	0.9225
	S4,L2	2.0,0.6	2.0057	0.5000	–	0.8551	2.0316	0.5000	–	0.9133
	S1,S3,S5	1.0,2.4,0.6	0.9061	0.9638	0.2871	0.3704	0.0601	1.0000	0.9415	0.9415
	S2,S4,S8	2.0,1.6,0.24	0.0972	0.9602	0.8467	0.8467	1.8074	0.6667	0.0000	0.1356
3	S6,S7,S9	1.0,0.8,0.36	1.1136	0.6454	0.0000	0.8994	4.9877	0.0010	0.0000	0.9937
	S1,L2,L4	1.0,0.8,0.6	2.0684	0.3333	–	0.7796	2.7640	0.3333	–	0.6913
	S7,S8,L3	1.0,0.32,0.6	2.2657	0.4309	–	0.3070	4.3079	0.0000	–	0.3980
	S1,S2,S4,S5	1.0,1.6,1.2,0.4	1.9804	0.5155	0.0011	0.0700	0.0341	0.9997	0.9645	0.9645
	S4,S6,S7,S8	2.0,0.8,0.6,0.16	0.7578	0.6303	0.0000	0.3558	3.0949	0.2500	0.0000	0.3854
4	S8,S5,S3,S1	0.4,0.8,1.8,0.4	1.2439	0.5336	0.0426	0.4889	1.4992	0.5124	0.0000	0.5210
	S8,S6,L4,L2	0.4,0.8,0.6,0.4	2.6502	0.0940	–	0.0034	3.2625	0.0000	–	0.0556
	S1,L2,L3,L4	1.0,0.8,0.6,0.4	2.3665	0.2500	–	0.3409	2.7862	0.2500	–	0.7017

Note: S_n means a bias in stream S_n and L_n a leak in unit n.

TABLE III Performance comparison between SEM and PCMT–SEM

Gross error introduced			*SEM*				*PCMT–SEM*			
No.	*Location*	*Size*	*AVTI*	*OP*	*OPF*	*OPFE*	*AVTI*	*OP*	*OPF*	*OPFE*
1	S15	500000	0.115	1.000	0.889	0.889	0.088	1.000	0.916	0.916
2	S21	60000	0.114	0.992	0.891	0.891	0.090	1.000	0.915	0.915
3	S15 S21	500000 60000	0.120	0.996	0.885	0.885	0.086	1.000	0.918	0.918
4	S14 S15 S23	−500000 −500000 50000	0.114	0.999	0.889	0.889	2.245	0.667	0.000	0.001
5	S4 S15 S37	7000000 500000 200000	0.115	0.999	0.890	0.890	0.083	1.000	0.922	0.922
6	S4 S15 S21 S37	700000 500000 60000 200000	0.187	0.994	0.827	0.827	0.081	1.000	0.924	0.924
7	S14 S15 S23 S56	−500000 −500000 50000 100000	0.329	0.892	0.552	0.565	4.235	0.500	0.000	0.000

Remark 3 It must be made clear that Tong and Crowe (1995) stated that the elements not retained by the principal component test can be picked up by a collective statistic. Incorporation of such strategy can change the above presented results. However, in the same fashion, the measurement test can also be aided by a global test to improve the detection, in a similar way as it was proposed by Sánchez *et al.* (1999). *Thus, the fair comparison would be to implement a serial elimination strategy that at each step would make use of PCMT and the collective test Q_e, exactly as suggested by Tong and Crowe* (1995), *or similarly, to determine which measurement should be eliminated and compare it with a strategy where the MT is used in conjunction with the global test. Such comparison may render a result that is favorable to the principal component strategy.*

CONCLUSIONS

The performance of the serial elimination strategy based on the measurement test and the principal component measurement test has been compared in this paper. The simulation results for a small example and a large size industrial system show that the use of PC tests does not necessarily improve significantly the power of serial identification strategies. In fact, it sometimes performs better and sometimes worse. For this reason, it appears that the performance of these methods is dependent on the location and size of the gross errors, a fact that the literature is already aware. In addition, there are cases, even for one gross error, where the hypothesis that the major contribution to the largest principal component fails to identify the correct set of gross errors. *All this should not be interpreted as a failure of principal component tests, but rather their inability to outperform the measurement test in the context described in this paper. Other implementations of serial elimination with PCMT aided by other tests could be successful.*

Acknowledgment

Partial financial support from KBC Advanced Technologies, now OSI, for Q. Jiang is acknowledged.

NOTATION

a	vector of measurement adjustments
A	$(m \times n)$ balance matrix

AVTI	Average number of Type I errors
g	vector of contributions to a suspect principal component
d	the size of a measurement group
$d_{\max}$	maximum number of measurements that can be deleted
$F_{\min}$	the desired degree of confidence
I	identity matrix
k_1	number of major contributors to a suspect principal component
K	matrix reflecting the positions of leaks
L	matrix reflecting the positions of biases
m	number of equations
n	number of measurements
npr	number of elements of p^r
OP	Overall power
OPF	expected fraction of correct identification
OPFE	expected fraction of successful identification
P	general distribution
p^r	principal component vector of vector r
p^a	principal component vector of vector a
Q	Collective test
r	equations' residuals
S	set of all measured stream flow rates
S_c	current set suspected of containing gross errors
S_d	temporary set being deleted in a particular step
t	number of non-zero eigenvalues of V
U_r	matrix of orthonormalized eigenvectors of Φ
U_a	matrix of orthonormalized eigenvectors of V
V	covariance matrix of a
W_r	matrix defined by Eq. (5)
W_a	matrix defined by Eq. (11)
y	vector of measurements

Greek Symbols

Ψ	measurement error covariance matrix
Φ	residual covariance matrix
Λ_r	diagonal matrix of the eigenvalues of Φ
Λ_a	diagonal matrix of the eigenvalues of V
δ	$(n \times 1)$ measurement biases
γ	$(m \times 1)$ leaks
ε_D, ε_E	threshold tolerances

τ_c critical value for the test statistic
ξ prescribed tolerance
$\lambda_{a,i}$ ith eigenvalue of V

References

Bagajewicz, M. and Jiang, Q. (1998) Gross Error Modeling and Detection in Plant Linear Dynamic Reconciliation. *Computers and Chemical Engineering*, **22**(12), 1789–1810.

Crowe, C. M., García Campos, Y. A. and Hrymak, A. (1983) Reconciliation of Process Flow Rates by Matrix Projection Part I: Linear Case, *AIChE J.*, **29**, 881–888.

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

Iordache, C., Mah, R. and Tamhane, A. (1985) Performance Studies of the Measurement Test for Detection of Gross Errors in Process Data. *AIChE J.*, **31**, 1187–1201.

Jiang, Q. and Bagajewicz, M. (1999) On a Strategy of Serial Identification with Collective Compensation for Multiple Gross Error Estimation in Linear Steady State Reconciliation. *I & ECR*, **38**(5), 2119–2128.

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

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

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

Narasimhan, S. and Mah, R. (1987) Generalized Likelihood Ratio Method for Gross Error Identification, *AIChE J.*, **33**, 1514–1521.

Nogita, S. (1972) Statistical Test and Adjustment of Process Data, *Ind. Eng. Chem. Process Des. Dev.*, **2**, 197–200.

Ripps, D. L. (1965) Adjustment of Experimental Data, *Chem. Eng. Prog. Symp. Ser.*, **61**, 8–13.

Romagnoli, J. A. and Stephanopoulos, G. (1981) Rectification of Process Measurement Data in the Presence of Gross Errors, *Chem. Engng. Sci.*, **36**, 1849–1863.

Sánchez, M. and Romagnoli, J. (1996) Use of Orthogonal Transformations in Classification/ Data Reconciliation. *Comp. Chem. Engng.*, **20**, 483–493.

Sánchez, M., Romagnoli, J., Jiang, Q. and Bagajewicz, M. (1999) Simultaneous Estimations of Biases and Leaks in Process Plants. *Computers and Chemical Engineering*, **23**(7), 841–858.

Serth, R. and Heenan, W. (1986) Gross Error Detection and Data Reconciliation in Steam Metering Systems. *AIChE J.*, **32**, 733–742.

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

Tong, H. and Crowe, C. (1995) Detection of Gross Errors in Data Reconciliation by Principal Component Analysis, *AIChE J.*, **41**(7), 1712–1722.

Tong, H. and Crowe, C. (1996) Detecting Persistent Gross Errors by Sequential Analysis of Principal Components. *Comp. Chem. Engng.*, **S20**, S733–S738.

Tong, H. and Bluck, D. (1998) An Industrial Application of Principal Component Test to Fault Detection and Identification. *IFAC Workshop on On-Line-Fault Detection and Supervision in the Chemical Process Industries*, Solaize (Lyon), France.

Václavek, V. (1969) Studies on System Engineering III. Optimal Choice of the Balance Measurements in complicated Chemical Engineering Systems. *Chem. Eng. Sci.*, **24**, 947–955.

APPENDIX

Streams and Measurements in the Second Example

No.	*Stream*	*From*	*To*	*Measurement*
1.	S1	U1	U2	691896
2.	S2	U1	U3	546045
3.	S3	U1	U4	777469
4.	S4	U5	U6	13247819
5.	S5	U3	U7	2474004
6.	S6	U8	ENV	1474762
7.	S7	U8	U9	600126
8.	S8	U8	U10	2554794
9.	S9	U8	U11	1653102
10.	S10	U8	U12	480902
11.	S11	U13	U5	35647542
12.	S12	U14	U15	11327255
13.	S13	U11	U14	11030529
14.	S14	U15	U16	8725778
15.	S15	U5	U11	8717214
16.	S16	U10	ENV	202400
17.	S17	U17	U10	6963023
18.	S18	U10	U9	707327
19.	S19	U5	U18	2369681
20.	S20	U10	U19	5911092
21.	S21	U18	U1	1047192
22.	S22	U10	U11	679275
23.	S23	U18	U4	952437
24.	S24	U4	U20	1715219
25.	S25	U10	U21	2303020
26.	S26	U12	U22	11164402
27.	S27	U5	U12	10806052
28.	S28	U9	U1	947308
29.	S29	U23	U24	6525339
30.	S30	U22	U23	11403839
31.	S31	U18	U3	1863497
32.	S32	ENV	U10	202400
33.	S33	U19	U25	1039951
34.	S34	U19	U26	4809074
35.	S35	U6	U17	6596938
36.	S36	U6	U8	6757151
37.	S37	U23	U18	4171828
38.	S38	U23	U27	654228
39.	S39	U27	U17	331622
40.	S40	U27	U14	332644
41.	S41	U23	U22	220414
42.	S42	U28	ENV	2831592
43.	S43	U18	U28	2502891
44.	S44	U9	U28	379713
45.	S45	U29	U13	1564750
46.	S46	ENV	U29	–
47.	S47	U30	U13	33101500
48.	S48	U31	U13	981300

No.	*Stream*	*From*	*To*	*Measurement*
49.	S49	ENV	U31	–
50.	S50	ENV	U32	–
51.	S51	U20	ENV	2792330
52.	S52	U7	ENV	760549
53.	S53	U24	ENV	11190400
54.	S54	U2	ENV	1269390
55.	S55	U33	ENV	3723350
56.	S56	U16	ENV	12839200
57.	S57	U21	ENV	5615070
58.	S58	U26	ENV	7510800
59.	S59	U25	ENV	2447520
60.	S60	ENV	U34	1039740
61.	S61	ENV	U30	17150378
62.	S62	U29	U30	17150378
63.	S63	U32	U30	17150400
64.	S64	ENV	U32	0
65.	S65	ENV	U30	17150378
66.	S66	ENV	U21	9092789.85
67.	S67	ENV	U29	13617185.88
68.	S68	ENV	U7	4212909.84
69.	S69	ENV	U2	2343106.24
70.	S70	ENV	U30	10416883.49
71.	S71	ENV	U24	15299608.42
72.	S72	ENV	U32	13058224.37
73.	S73	ENV	U31	22078202.75
74.	S74	ENV	U20	4442389.94
75.	S75	ENV	U26	7523307.20
76.	S76	ENV	U25	2963668.35
77.	S77	ENV	U33	2522343.23
78.	S78	ENV	U34	2396928.73
79.	S79	ENV	U16	9380521.93
80.	S80	U21	ENV	5786320.81
81.	S81	U29	ENV	11434279.89
82.	S82	U7	ENV	5904762.80
83.	S83	U35	ENV	1762218.53
84.	S84	U30	ENV	10416883.49
85.	S85	U24	ENV	10753445.20
86.	S86	U32	ENV	17887554.39
87.	S87	U31	ENV	12642438.49
88.	S88	U20	ENV	3376982.72
89.	S89	U26	ENV	4868019.18
90.	S90	U25	ENV	1557652.15
91.	S91	U33	ENV	1494720.98
92.	S92	U34	ENV	3437115.69
93.	S93	U16	ENV	5386164.63

Note: ENV represents the environmental node.