On the Performance of Principal Component Analysis in Multiple Gross Error Identification

Qiyou Jiang, Mabel Sánchez,† and Miguel Bagajewicz*

School of Chemical Engineering and Materials Science, University of Oklahoma, The Energy Center, 100 E. Boyd, Room T-335, Norman, Oklahoma 73019-0628, and Planta Piloto de Ingeniería Química (UNS-CONICET), Camino La Carrindanga, Km 7, (8000) Bahía Blanca, Argentina

In this paper, the use of the principal component test for the identification stage of three existing collective compensation strategies is presented. The three modified techniques are UBET (unbiased estimation of gross errors), SEGE (simultaneous estimation of gross errors) in the form of their recent modifications (MUBET and MSEGE), and SICC (serial identification with collective compensation). These techniques are modified to apply a statistical test based on principal component analysis instead of the nodal, global, and measurement tests they use. The performance of the modified techniques is assessed by means of Monte Carlo simulations. Comparative analysis indicates that PCA tests do not significantly enhance the ability in identification features of these strategies, and even in some cases, it may lower the exact identification performance.

Introduction

Typically, the presence of biased instruments and leaks, as well as departures from steady state, invalidates the results of data reconciliation techniques. These 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. The application of a hypothesis statistical test has been extensively used for this purpose. The global test,[1] the measurement test,[2,3] the nodal test,[1,4] the generalized likelihood ratio (GLR)[5] test, and Bonferroni tests[6] have been used for gross error identification. Recently, principal component tests (PCT) have been proposed by Tong and Crowe.[7,8] Some results on their performance were reported by Tong and Bluck.[9] The authors indicated that tests based on principal component analysis (PCA) are more sensitive to subtle gross errors and have greater power to correctly identify the variables in error than the first three tests.

For multiple gross error identification and estimation, serial elimination, serial compensation, and simultaneous or collective compensation strategies have been proposed. Among them, collective compensation strategies have some advantages: they are applicable to all types of gross errors, can maintain redundancy during the procedure, and provide better estimates thanks to the collective estimation.[6]

Collective compensation strategies have been presented for steady-state linear processes by Rollins and Davis,[6] Keller et al.,[10] Kim et al.,[11] and Sánchez and Romagnoli,[12] that considered the simultaneous estimation of all gross errors. In addition, Bagajewicz and Jiang[13] proposed a collective compensation strategy for dynamic systems that can be used for steady-state cases. The methods presented by Keller et al.,[10] Rollins and Davis,[6] and Sánchez and Romagnoli[12] (CGLR, UBET, and SEGE) have been recently modified to avoid singularities and to assess uncertainties (Bagajewicz et al.[14] and Sánchez et al.[15]).

The increasing application of PCA in process monitoring and fault diagnosis (Gertler et al.,[16] Tong and Bluck,[9] Vedam and Venkatasubramanian,[17] and Jia et al.[18]) suggests the need of reviewing these methods in the area of gross error identification and comparing their performance. In this work, the performance of PCA tests will be evaluated in a systematic way using Monte Carlo simulation (Iordache et al.[19]). For this analysis, the statistical tests applied in the identification step of three existing collective compensation strategies are replaced with PCA tests, while the estimation stage remains unchanged. The study considers the following techniques: UBET, unbiased estimation of gross errors[6] as modified by Bagajewicz et al.;[14] SEGE, simultaneous estimation of gross errors[12] as modified by Sánchez et al.;[15] and SICC, serial identification with collective compensation,[20] for which a variant is also proposed in this paper.

The paper is organized as follows: The PCT is reviewed first. The modified versions of UBET and SEGE (MUBET and MSEGE) are reviewed next, followed by a discussion of modifications performed to SICC. Last, results are discussed.

Principal Component Tests

In this section a review of principal component tests proposed by Tong and Crowe[7] is presented. Both 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

$$\mathbf{r} = \mathbf{A}\mathbf{y} \tag{1}$$

where $\mathbf{y}$ is the vector of measurements and $\mathbf{A}$ the

* To whom correspondence should be addressed. Phone: (405) 325-5458. Fax: (405) 325-5813. E-mail: bagajewicz@ou.edu.

† Phone: 54-91-861700. Fax: 54-91-861600. E-mail: msanchez@plapiqui.edu.ar.

10.1021/ie9804677 CCC: $18.00
Published on Web 03/24/1999

balance matrix. For simplicity, in this formulation it is assumed that all variables are measured. This is not a limitation because the unmeasured variables can be removed using, for example, the reduced balance scheme,[21] matrix projection,[3] QR decomposition,[22,23] or matrix co-optation.[24]

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

$$E\{\mathbf{r}\} = \mathbf{r}^* = 0 \qquad \Phi = \mathrm{cov}\{\mathbf{r}\} = \mathbf{A}\Psi\mathbf{A}^{\mathrm{T}} \tag{2}$$

The eigenvalue decomposition of matrix Φ is formulated as

$$\mathbf{A}_\mathbf{r} = \mathbf{U}_\mathbf{r}{}^{\mathrm{T}}\Phi\mathbf{U}_\mathbf{r} \tag{3}$$

where $\mathbf{U}_\mathbf{r}$ is the matrix of orthonormalized eigenvectors of Φ ($\mathbf{U}_\mathbf{r}\mathbf{U}_\mathbf{r}{}^{\mathrm{T}} = \mathbf{I}$) and $\Lambda_\mathbf{r}$ is the diagonal matrix of the eigenvalues of Φ.

Accordingly, the following linear combinations of r are proposed

$$\mathbf{P}^\mathbf{r} = \mathbf{W}_\mathbf{r}{}^{\mathrm{T}}(\mathbf{r} - \mathbf{r}^*) = \mathbf{W}_\mathbf{r}{}^{\mathrm{T}}\mathbf{r} \tag{4}$$

where

$$\mathbf{W}_\mathbf{r} = \mathbf{U}_\mathbf{r}\Lambda_\mathbf{r}{}^{-1/2} \tag{5}$$

and $\mathbf{p}^\mathbf{r}$ consists of the principal components. Also, $r \sim P(0,\Phi) \Rightarrow \mathbf{p}^\mathbf{r} \sim P(0,\mathbf{I})$, for any distribution P. That is, a set of correlated variables r is transformed into a new set of uncorrelated variables $\mathbf{p}^\mathbf{r}$. If the measurement errors are normally distributed, then the principal components will be normally distributed too. That is, $\mathbf{y} \sim N(x,\Psi) \Rightarrow \mathbf{p}^\mathbf{r} \sim N(0,\mathbf{I})$. Consequently, instead of looking at a statistical test for r, the hypothesis test may be performed on $\mathbf{p}^\mathbf{r}$. Tong and Crowe[7] proposed the following principal component nodal test:

$$\mathbf{P}_i{}^\mathbf{r} = (\mathbf{W}\mathbf{r}^{\mathrm{T}}\mathbf{r})_i \sim N(0,1) \qquad i = 1, ..., \mathrm{npr} \tag{6}$$

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 $p_i{}^\mathbf{r}$, which can be calculated by

$$g_j = (w_i{}^\mathbf{r})_j r_j \qquad j = 1, ..., m \tag{7}$$

where $w_i{}^\mathbf{r}$ is the ith eigenvector in $\mathbf{W}_\mathbf{r}$.

Similarly, the principal component measurement test can be stated as follows:

The vector of adjustments $\mathbf{a}$ and its covariance matrix $\mathbf{V}$ are

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

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

The vector of principal components, $\mathbf{p}^\mathbf{a}$, is the following linear combination of $\mathbf{a}$:

$$\mathbf{p}^\mathbf{a} = \mathbf{W}_\mathbf{a}{}^{\mathrm{T}}(\mathbf{a} - \mathbf{a}^*) = \mathbf{W}_\mathbf{d}{}^{\mathrm{T}}\mathbf{a} \tag{10}$$

Because $\mathbf{V}$ is singular, $\Lambda_\mathbf{a}$ has some eigenvalues which are zero. Then, instead of normalizing the whole vector $\mathbf{p}^\mathbf{a}$ and applying 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

$$w_{\mathbf{a},i} = \mathbf{U}_{\mathbf{a},i}\lambda_{\mathbf{a},i}{}^{-1/2} \qquad i = 1, ..., t \tag{11}$$

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

$$\Lambda_\mathbf{a} = \mathbf{U}_\mathbf{a}{}^{\mathrm{T}}\mathbf{V}\mathbf{U}_\mathbf{a} \tag{12}$$

Thus, the normalized principal components are

$$p_i{}^\mathbf{a} = (w_i{}^\mathbf{a})^{\mathrm{T}}\mathbf{a} \qquad i = 1, ..., t \tag{13}$$

Then, $\mathbf{a} \sim P(0,\mathbf{V}) \Rightarrow \mathbf{p}^\mathbf{a} \sim P(0,\mathbf{T})$. When the adjustments are normally distributed, Tong and Crowe[7] proposed the 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, $p_i{}^\mathbf{a}$. This contribution is calculated as follows:

$$g_j = (w_i{}^\mathbf{a})_j a_j \qquad j = 1, ..., n \tag{14}$$

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

$$\left| \frac{\sum^{k_1} g'_j - p_i{}^\mathbf{a}}{p_i{}^\mathbf{a}} \right| \le \zeta \tag{15}$$

where ζ may be fixed, for example, at 0.1.

Collective Compensation Strategies

In this section, three collective compensation strategies are reviewed.

(a) Review of MUBET. The unbiased estimation technique, UBET,[6] is developed from the balance residuals, defined by eq 1 and its expected value

$$\mu_\mathbf{r} = \mathbf{A}\delta + \mathbf{M}\gamma \tag{16}$$

where

$$\mathbf{M} = [m_1, ..., m_q] \tag{17}$$

and n and q are the number of measured variables and constraint equations, respectively, δ and γ are the unknown $n \times 1$ measurement biases and $q \times 1$ leaks, $\mathbf{A}$ is a $q \times n$ constraint matrix with rank ($\mathbf{A}$) = q, and m_j is a $q \times 1$ vector with zeros in every position but a 1 in the jth.

By partitioning **A**, **M**, δ, and γ and assuming there are always q gross errors, one can get

$$\mu_{\mathbf{r}} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \gamma_2 \end{bmatrix} = \mathbf{C}_1 \theta_1 \tag{18}$$

Finally, by introducing

$$\mathbf{l}_i^{\mathrm{T}} = \mathbf{e}_i^{\mathrm{T}} \mathbf{C}_1^{-1} \tag{19}$$

one obtains

$$\mathbf{l}_i^{\mathrm{T}} \mu_{\mathbf{r}} = \mathbf{e}_i^{\mathrm{T}} \theta_1 = \theta_i \tag{20}$$

Thus, $\mathbf{l}_i^{\mathrm{T}}\mathbf{r}$ ($i = 1, ..., q$) are unbiased estimators of the components of δ and γ contained in θ_1.

According to Rollins and Davis,[6] the procedure for applying UBET can be summarized as follows:

(1) Use a gross error identification strategy, like the nodal strategies reported by Mah et al.[4] and Serth and Heenan,[25] to isolate the suspect nodes and construct the candidate bias/leak list from the suspect nodes.

(2) Obtain θ_1 with elements no more than the number of constraint equations (q).

(3) Construct $\mathbf{C}_1$ with rank equal to q.

(4) Obtain the size estimation for the elements in θ_1.

(5) Use the Bonferroni test to identify the gross errors.

Bagajewicz et al.[14] have presented a modified version of this strategy, which addresses singularities and uncertainties of the original method (MUBET). The modifications are the following:

(i) In step 1, after the candidate bias/leak list is constructed, the equivalency theory[13] is applied to delete the possible existing loops in this list. All candidates are checked one by one. If a candidate forms a loop with any candidate(s) checked, it will be deleted from the list.

(ii) After this modified step 1, in view of the equivalency theory, the elements of θ_1 in step 2 will certainly be no more than the number of constraint equations (q). Therefore, θ_1 can be directly transferred from step 1.

(iii) In step 3, if the number of elements in θ_1 is less than q, one has to add other streams/leaks as candidates to θ_1 to make the number equal to q. Any of these streams/leaks must be checked to be sure that it is not forming a loop with any element(s) in θ_1 before it is added.

(b) Review of MSEGE. In this technique, a collective statistical test based on the vector of adjustments a is selected to detect the presence of gross errors. The null and alternative hypotheses are stated as follows: $H_0 = \mathbf{a}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{a}$, $E(\mathbf{a}) = 0$; $H_1 = E(\mathbf{a}) \neq 0$. The global test statistic H_0 is used to compare both alternatives.

If H_0 is rejected, the first stage of the procedure starts. This allows the isolation of a subset of constraints that do not pass the global test. This applies a recursive procedure that has the advantage that only the reciprocal of a scalar has to be computed in each step. In this procedure, equations are added one by one to the least-squares estimation problem of the vector **x**. After each addition, the objective function (ofv) of the least-squares estimation technique is calculated and compared with the critical value τ_c. For updating the test statistic, the following expressions are applied

$$\Sigma_c^{\mathrm{new}} = \Sigma_c^{\mathrm{old}} - \Sigma_c^{\mathrm{old}} \mathbf{B}_i^{\mathrm{T}} (\mathbf{B}_i \Sigma_c^{\mathrm{old}} \mathbf{B}_i^{\mathrm{T}})^{-1} \mathbf{B}_i \Sigma_c^{\mathrm{old}} \tag{21}$$

$$\hat{x}^{\mathrm{new}} = \Sigma_c^{\mathrm{new}} \Psi^{-1} \mathbf{y} \tag{22}$$

$$\mathrm{ofv} = (\mathbf{y} - \hat{x}) \Psi^{-1} (\mathbf{y} - \hat{x}) \tag{23}$$

where Σ_c^{new} and Σ_c^{old} represent the covariance matrices of the measurement estimates after and before equation addition, and $\mathbf{B}_i$ stands for the added equation.

After an equation is incorporated, the following is checked:

(a) If ofv $> \tau_c$, gross errors are detected, so the last equation of the system of equations is eliminated. All of the measurements involved in the constraint and a leak from the corresponding node are added to the list of suspected gross errors.

(b) If ofv $< \tau_c$, gross errors are not detected after the addition of the constraint, and the constraint remains in the set.

From stage 1 of the procedure, a set of measurements and units suspected of being biased or having leaks is obtained. In stage 2, the identification and estimation of gross errors is accomplished by the following procedure:

(1) Set the number of gross errors s to 1 ($s = 1$).

(2) Take all combinations of s gross errors and run the corresponding reconciliation model (biases only, leaks only, or biases and leaks).

(3) Determine which combination of gross errors gives the lowest objective function value.

(4) If the global test is satisfied, stop. The combination or combinations with the lowest objective function value involve gross errors. Otherwise, increase s by 1 ($s = s + 1$) and go to step 2.

Sánchez et al.[15] have modified this strategy to address singularities and uncertainties of the original method (MSEGE). The modifications are as follows:

(i) If a set to be investigated in stage 2 is equivalent to a set previously considered, the set is ignored, as the result is known.

(ii) If a set to be investigated in stage 2 is included in any loop of the augmented graph, the set is ignored. It is known that it will be singular, as it cannot capture the number of gross errors targeted.

(iii) All equivalent sets are identified at the end of the application of the algorithm.

(C) Review of SICC. This strategy relies on the measurement test for gross error detection. It uses the MT to make a list of suspect gross errors and identifies from the list one gross error using a compensation model.[13] This error is put in a list of confirmed gross errors. Next a new list of suspects is constructed, and the compensation model is run using the confirmed gross errors and a new candidate at a time to determine which should be added to the confirmed gross error list. The procedure is repeated until no gross errors are detected. Leaks are identified using the equivalency theory.

The technique is made up of the following steps:

(1) Run the data reconciliation and calculate the measurement tests (MT). If there are no MT flags, declare no gross error and stop. Otherwise, go to step 2.

(2) Construct a list of candidates (LC) by including all variables that failed the MT. If any two members in

LC form a loop, erase one of them. Create a list of confirmed gross errors (LCGE). This list is empty at this stage.

(3) Run the data reconciliation with the gross error estimation model simulating a gross error in all of the members of the LCGE and in one member of the LC at a time.

(4) Determine which member of the LC leads to the smallest value of the objective function. Add that variable to the LCGE.

(5) Calculate MT for the run chosen in step 4. Erase all elements of the LC and place the latest flagged variables in LC. If there are any two members in LC forming a loop with any member(s) in LCGE, erase one of them. If LC is empty, go to step 6. Otherwise, go to step 3.

(6) Determine all equivalent sets and corresponding gross error sizes. Declare all members in LCGE in suspect and stop.

A Nodal Test Based SICC (NT-SICC)

The original version of SICC relies on the measurement test for the identification step. To identify leaks, this method needs to resort to equivalency theory, as the compensation model only uses biases. The nodal test can be used to construct a list of candidates. This list will now include leaks, and thus the full compensation module with biases and leaks can be used. The modifications are summarized next:

(i) In step 1: Calculate the nodal tests (NT). If there are no NT flags, declare no gross error and stop. Otherwise, go to step 2.

(ii) In step 2: Construct a list of candidates (LC) by including all streams and leaks connected to the nodal tests that have failed. If any two members in LC form a loop, erase one of them. Create a list of confirmed gross errors (LCGE). This list is empty at this stage.

(iii) In step 5: Calculate NT for the run chosen in step 4. Erase all elements of the LC and place the latest flagged variables in LC. If any two members in LC form a loop, erase one of them. If there is any member in LC forming a loop with any member in LCGE, erase it from LC. If there are any two members in LC forming a loop with any member in LCGE, erase one of them. If LC is empty, go to step 6. Otherwise, go to step 3.

Inclusion of Principal Component Tests

The aforementioned techniques have been modified to use PCA tests. For MUBET, the strategy of pseudonodes[25] has been replaced by PCNT. However, the Bonferroni tests used at the end of the estimation step have been left unaltered. The same modification is performed to MSEGE: Its stage 1 is accomplished now by PCNT to obtain a suspect set of measurements and leaks. In the case of SICC, the measurement test was replaced by the PCMT test. Finally, in the case of the nodal test-based SICC (NT-SICC), the nodal test was replaced by PCNT.

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.[19] was followed. Each result is based on 10 000 simulation trials where the random errors are changed and the magnitudes of gross errors are fixed.

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

$$\mathrm{OP} = \frac{\text{no. of gross errors correctly identified}}{\text{no. of gross errors simulated}} \tag{24}$$

$$\mathrm{AVTI} = \frac{\text{no. of gross errors incorrectly identified}}{\text{no. of simulation trials}} \tag{25}$$

$$\mathrm{OPF} = \frac{\text{no. of trials with perfect identification}}{\text{no. of simulation trials}} \tag{26}$$

The first two measures are proposed by Mah and Narasimhan[5] and the last one by Rollins and Davis.[6]

In addition, a set of gross errors may have its equivalent basic sets, as described by Bagajewicz and Jiang.[13] Thus, to assess these uncertainties, a new measure, the overall performance of equivalent identification (OPFE), was introduced recently by Sánchez et al.[15]

$$\mathrm{OPFE} = \frac{\text{no. of trials with successful identification}}{\text{no. of simulation trials}} \tag{27}$$

(a) Determination of OPFE. To clarify the uncertainty in gross error detection, a series of concepts were presented in a recent paper.[13] 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:[20]

$$[\mathbf{AL}_1 \quad \mathbf{K}_1]\begin{bmatrix}\hat{\delta}_1\\ \hat{\gamma}_1\end{bmatrix} = [\mathbf{AL}_2 \quad \mathbf{K}_2]\begin{bmatrix}\hat{\delta}_2\\ \hat{\gamma}_2\end{bmatrix} \tag{28}$$

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

Premultiplying both $[\mathbf{AL}_1 \quad \mathbf{K}_1]$ and $[\mathbf{AL}_2 \quad \mathbf{K}_2]$ by a certain particular matrix, one can transform $[\mathbf{AL}_2 \quad \mathbf{K}_2]$ 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.[20] 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 $\delta_4 = +1$ and $\delta_5 = +1$. These variables are part of the equivalent set (S_2, S_4, and S_5), which has gross error cardinality 2. To determine whether the identification

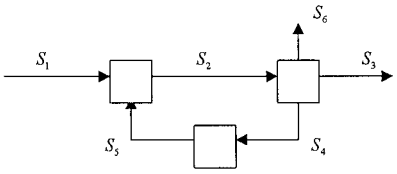

Figure 1. Example of a flowsheet.

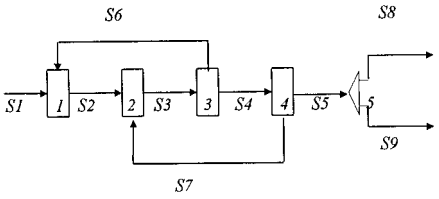

Figure 2. Example of a process flowsheet.

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

In eq 28, 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 τ_D.

Furthermore, another situation called quasi-equivalency of gross errors can arise. 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 $\delta_1 = +0.98$ and $\delta_2 = +0.05$. This is a type I error but is accompanied with a small size estimate. In principle, even though 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 by using eq 28 and a threshold tolerance τ_E.

Thus, OPFE measure represents the fraction of trials with successful identification. A trial is considered successful if the set of simulated gross errors is identified in the simulated positions or if the strategy identifies a set of gross errors which is equivalent to the simulated set. OPFE is calculated in this paper by allowing both quasi-degeneracy and quasi-equivalency.

(b) Differences in Success Measures. The performance evaluation of gross error identification strategies is intended to analyze their behavior in different situations (location and size of gross errors). In this paper the evaluation procedure, which is based on Monte Carlo simulations, proposes the use of three measures of success: OP, OPF, and OPFE. The main differences between them are the following:

(a) OP indicates the fraction of simulated gross errors that are identified. Thus, some trials may identify other errors that are neither simulated nor equivalent sets.

(b) OPF counts as successful those trials for which all of the simulated gross errors are identified. As a measure of exact identification, it is superior to OP because it takes into account the fact that no extra gross errors, except the simulated ones, are identified.

(c) OPFE is an extension of OPF, which avoids considering a failure of the algorithm when a set equivalent to the simulated set is identified. However, no other gross errors, except the equivalent sets, are allowed to participate in the identified set.

Results

First, 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 10 random generated values. To compare results on the same basis, the level of significance of each method was chosen such that it gives an AVTI equal to 0.1 under the null hypothesis. This common basis of comparison for gross error detection schemes was introduced by Rosenberg et al.[26] The same confidence levels for comparison have been used by different authors too. If a comparison is performed on such an alternative basis, the conclusions of this paper do not vary. Each strategy was tested under the same scenarios of gross errors introduced. The size of the gross error is selected as 4 times its corresponding flow rate standard deviation when there is only one gross error, 5 and 3 times for two gross errors, and 5, 4, and 3 times for three gross errors. If a leak is introduced, then the minimum standard deviation of flow rates 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 the claim has been that PCA is more efficient in the case of small gross errors. For this reason, the thresholds τ_D and τ_E are quite permissive. They both have been chosen as the maximum gross error size in the introduced set. If a more stringent criterion is used, the values of OPFE may be lower.

Table 1 shows the comparison between MUBET and PCNT-MUBET. When there is one gross error, both of them reached high and similar performance, i.e., high OP, OPF, and OPFE with low AVTI. For more than one gross error, the OPF for the two methods drops to low values, reaching zero for the case of three gross errors. However, the OPFE maintains good values and for PCNT-MUBET is slightly higher in most cases.

The AVTI of all of 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 consideration of the criterion of perfect identification.

Table 1. Performance Comparison between MUBET and PCNT-MUBET

gross error introduced			MUBET				PCNT-MUBET			
no.	location[a]	size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S_1	0.8	0.1582	0.9626	0.8793	0.9978	0.1338	1.0000	0.8809	0.9992
	S_3	2.4	0.1141	0.9927	0.9066	0.9998	0.1717	1.0000	0.8414	0.9992
	S_5	0.8	0.2290	0.9431	0.8898	0.9984	0.1731	0.9995	0.8645	0.9991
	S_7	0.8	0.4162	0.8725	0.8165	0.9753	0.1590	0.9575	0.8234	0.9577
	L_3	0.8	0.1769	0.8337	0.7835	0.8756	1.7290	0.2983	0.2752	0.9678
2	S_1, S_4	1.0, 1.2	2.8005	0.0780	0.0686	0.9977	0.1189	0.9998	0.8896	0.9988
	S_3, S_8	3.0, 0.24	0.7924	0.5828	0.1641	0.8155	1.3226	0.6836	0.3360	0.8661
	S_6, S_7	1.0, 0.6	2.7848	0.2515	0.1285	0.6382	1.7704	0.5000	0.0000	0.7061
	S_2, S_5	2.0, 0.6	1.2664	0.7030	0.3883	0.9557	0.1357	0.9993	0.8762	0.9981
	S_4, L_2	2.0, 0.6	0.2287	0.8095	0.5876	0.6397	2.5140	0.5000	0.0000	0.7045
3	S_1, S_3, S_5	1.0, 2.4, 0.6	2.4214	0.4949	0.1769	0.9560	0.0986	0.9982	0.9127	0.9996
	S_2, S_4, S_8	2.0, 1.6, 0.24	4.2424	0.0969	0.0000	0.8155	1.9111	0.6663	0.0000	0.8608
	S_6, S_7, S_9	1.0, 0.8, 0.36	4.8456	0.0013	0.0000	0.8728	3.8903	0.3335	0.0000	0.9150
	S_1, L_2, L_4	1.0, 0.8, 0.6	1.8961	0.3733	0.0005	0.9454	2.1917	0.3333	0.0000	0.8997
	S_7, S_8, L_3	1.0, 0.32, 0.6	2.3781	0.3641	0.0007	0.6074	1.8516	0.6072	0.0000	0.6866

[a] S_n means a bias in stream S_n and L_n a leak in unit n.

Table 2. Performance Comparison between MSEGE and PCNT-MSEGE

gross error introduced			MSEGE				PCNT-MSEGE			
no.	location	size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S_1	0.8	0.0448	0.9950	0.9593	0.9987	0.1021	0.9928	0.9066	0.9984
	S_3	2.4	0.0499	0.9932	0.9578	1.0000	0.1091	1.0000	0.9035	0.9978
	S_5	0.8	0.0434	0.9998	0.9577	0.9999	0.1006	0.9999	0.9025	1.0000
	S_7	0.8	0.0606	0.9776	0.9425	0.9785	0.1147	0.9776	0.8917	0.9776
	L_3	0.8	0.0840	0.9555	0.9326	0.9682	0.1328	0.9548	0.8943	0.9712
2	S_1, S_4	1.0, 1.2	0.6047	0.7105	0.6940	0.9986	0.6383	0.7127	0.6679	0.9991
	S_3, S_8	3.0, 0.24	0.2165	0.8559	0.7038	0.7487	1.0820	0.7275	0.4189	0.6151
	S_6, S_7	1.0, 0.6	0.1335	0.9337	0.8481	0.8821	0.1992	0.9342	0.8219	0.9003
	S_2, S_5	2.0, 0.6	0.0475	0.9989	0.9551	0.9979	0.1052	0.9982	0.8998	0.9973
	S_4, L_2	2.0, 0.6	0.3767	0.8321	0.6520	0.6822	0.1791	0.9451	0.8507	0.9134
3	S_1, S_3, S_5	1.0, 2.4, 0.6	0.0655	0.9930	0.9540	0.9970	0.1044	0.9986	0.9014	0.9952
	S_2, S_4, S_8	2.0, 1.6, 0.24	0.1747	0.9166	0.7740	0.8225	1.7376	0.6572	0.0988	0.8838
	S_6, S_7, S_9	1.0, 0.8, 0.36	1.0947	0.6449	0.0000	0.9572	1.1736	0.6352	0.0000	0.9659
	S_1, L_2, L_4	1.0, 0.8, 0.6	1.9335	0.3444	0.0000	0.9151	2.0212	0.3386	0.0000	0.9472
	S_7, S_8, L_3	1.0, 0.32, 0.6	0.8449	0.7014	0.4061	0.6501	0.6601	0.7782	0.5466	0.9093

Table 3. Performance Comparison between SICC and PCMT-SICC

gross error introduced			SICC				PCMT-SICC			
no.	location	size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S_1	0.8	0.0713	1.0000	0.9293	0.9999	0.0653	1.0000	0.9353	1.0000
	S_3	2.4	0.0693	1.0000	0.9317	0.9999	0.0790	1.0000	0.9231	0.9996
	S_5	0.8	0.0731	1.0000	0.9274	1.0000	0.0631	1.0000	0.9372	1.0000
	S_7	0.8	0.0820	0.9787	0.9202	0.9797	1.1429	0.4236	0.4032	0.9473
	L_3	0.8	2.0355	0.0000	0.0000	0.9787	2.5566	0.0000	0.0000	0.9403
2	S_1, S_4	1.0, 1.2	0.0465	0.9988	0.9543	0.9983	0.0235	0.9995	0.9769	0.9994
	S_3, S_8	3.0, 0.24	0.0717	0.9499	0.8749	0.9213	0.9851	0.5000	0.0000	0.9183
	S_6, S_7	1.0, 0.6	0.1380	0.9438	0.8592	0.8983	1.3892	0.5091	0.0212	0.4978
	S_2, S_5	2.0, 0.6	0.0478	1.0000	0.9523	1.0000	0.0241	1.0000	0.9760	1.0000
	S_4, L_2	2.0, 0.6	1.9792	0.5000	0.0000	0.8790	1.8590	0.5000	0.0000	0.7987
3	S_1, S_3, S_5	1.0, 2.4, 0.6	0.9413	0.9244	0.2985	0.9360	1.6100	0.7868	0.2969	0.8284
	S_2, S_4, S_8	2.0, 1.6, 0.24	0.0685	0.9530	0.8513	0.8882	0.9486	0.6667	0.0000	0.9049
	S_6, S_7, S_9	1.0, 0.8, 0.36	1.0685	0.6455	0.0000	0.9467	1.7177	0.6796	0.0460	0.8146
	S_1, L_2, L_4	1.0, 0.8, 0.6	2.0020	0.3333	0.0000	0.8150	2.0254	0.3333	0.0000	0.8798
	S_7, S_8, L_3	1.0, 0.32, 0.6	2.0572	0.3963	0.0000	0.4052	3.1281	0.1215	0.0000	0.6391

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.

A comparison between MSEGE and PCNT-MSEGE is shown in Table 2. For one gross error, MSEGE has lower AVTI and higher OPF while OP and OPFE are comparable to those of PCNT-MSEGE. For two gross errors, SEGE also has better OPF. When OPFE is considered, PCNT-MSEGE outperforms MSEGE 5 times out of 10 for two or three gross errors. Thus, in terms of perfect identification, PCA helps MSEGE strategy in two situations, also makes no difference in two situations, and performs worse in 11 cases.

Table 3 indicates the simulation results for SICC and PCMT-SICC. Performance measures show both strategies have similar performance. For instance, out of 15 cases, the OPFE is higher in 6 cases for SICC and higher in 3 cases for PCMT-SICC. The rest are about the same.

Table 4 shows a comparative performance of NT-SICC and PCNT-SICC. Although performance measures are slightly higher for PCNT-SICC, the conclusion that one should make is that the inclusion of PCT does not introduce any significant enhancement for the behavior of NT-SICC.

Table 4. Performance Comparison between NT-SICC and PCNT-SICC

gross error introduced			NT-SICC				PCNT-SICC			
no.	location	size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S_1	0.8	0.0850	0.9980	0.9156	0.9983	0.0668	0.9975	0.9345	0.9980
	S_3	2.4	0.0635	1.0000	0.9379	0.9998	0.0729	1.0000	0.9280	0.9999
	S_5	0.8	0.0655	1.0000	0.9354	0.9998	0.0659	1.0000	0.9344	0.9999
	S_7	0.8	0.0761	0.9784	0.9267	0.9800	0.0701	0.9780	0.9309	0.9790
	L_3	0.8	0.0997	0.8667	0.8015	0.8737	0.0945	0.9704	0.9088	0.9722
2	S_1, S_4	1.0, 1.2	1.4408	0.3518	0.2959	0.9993	1.4167	0.3321	0.3042	0.9986
	S_3, S_8	3.0, 0.24	0.0664	0.8842	0.7395	0.7920	0.1333	0.9042	0.7761	0.8372
	S_6, S_7	1.0, 0.6	0.1257	0.9363	0.8564	0.8847	0.1268	0.9258	0.8268	0.8639
	S_2, S_5	2.0, 0.6	0.0314	1.0000	0.9687	1.0000	0.0293	1.0000	0.9707	0.9999
	S_4, L_2	2.0, 0.6	0.3167	0.7550	0.4851	0.6105	0.3484	0.8830	0.7499	0.8590
3	S_1, S_3, S_5	1.0, 2.4, 0.6	1.1615	0.8955	0.1564	0.9654	1.1299	0.9070	0.1578	0.9768
	S_2, S_4, S_8	2.0, 1.6, 0.24	0.0379	0.9198	0.7567	0.7920	0.0988	0.9199	0.7564	0.7972
	S_6, S_7, S_9	1.0, 0.8, 0.36	1.1174	0.6340	0.0000	0.9435	1.1250	0.6335	0.0000	0.9414
	S_1, L_2, L_4	1.0, 0.8, 0.6	0.9648	0.6507	0.0000	0.8623	1.6588	0.4156	0.0000	0.8728
	S_7, S_8, L_3	1.0, 0.32, 0.6	1.1991	0.4742	0.1942	0.5753	1.1790	0.5896	0.2408	0.8430

These results clearly show that PCA is of help in some situations and it does not make a difference in others. Moreover, its application lowers the exact identification performance of some strategies that are analyzed in this paper. In this sense, PCA methods may be considered as members of the broad spectrum of gross error detection techniques and not superior strategies.

Although PCA strategies rely first on testing independent variables against the normal distribution, in the next steps the identification is broadened to the entire set of residuals or adjustments. The contributions to a suspect principal component are partially added to assess the number of major contributors k_1. There are some numerical evidences that this procedure may not be satisfactory enough because it enlarges the number of residuals/measurements in suspect, depending on the numerical values of the contributions and the threshold value.

It has been discussed elsewhere[13] that the use of OPF as an assessment tool for the power of gross error detection is misleading. The reason for this, it was argued, is the fact that equivalent sets have an effect on data reconciliation and therefore the identification of equivalent gross errors should be considered a success. However, for all practical purposes, high OPF may still be a desired property, as it reflects exact identification. However, as is shown from the examples, once one departs from simple cases of one gross error, this measure ceases to be practical. Thus, in the presence of several gross errors, the identification of equivalent sets is a must in practice. In addition, these equivalent sets represent an uncertainty that, in the absence of other information, is unavoidable.

Finally, if one chooses to use OPF instead of OPFE as a comparison tool, except for the fact that all techniques fail rather dramatically for a large number of gross errors when OPF is used, the conclusions about the introduction of PCA are the same.

Conclusions

The principal component tests have been added to different collective compensation techniques. The performance has then been compared to the regular techniques using known tests. The simulation results show that the use of PC tests does not necessarily improve the power of serial identification strategies. In fact, it sometimes performs better and sometimes worse. It also appears then that the performance of these methods is dependent on the location of the gross errors.

Acknowledgment

Partial financial support from KBC Advanced Technologies for Q.J. is acknowledged.

Notation

$\mathbf{a}$ = vector of measurement adjustments
$\mathbf{A}$ = $m \times n$ balance matrix
AVTI = average number of type I errors
$\mathbf{B}_i$ = row vector corresponding to a balance constraint
$\mathbf{C}_1$ = matrix defined by eq 18
$\mathbf{e}_i$ = vector with 1 in the ith place and zero elsewhere
$\mathbf{g}$ = vector of contributions to a suspect principal component
$\mathbf{I}$ = identity matrix
k_1 = number of major contributors to a suspect principal component
$\mathbf{K}$ = matrix for leaks' positions
$\mathbf{l}$ = vector used for making linear combinations of measurements
$\mathbf{L}$ = matrix for biases' positions
$\mathbf{M}$ = diagonal matrix of constants for calculating leaks
n = number of measurements
npr = number of elements of $p^{\mathbf{r}}$
npa = number of elements of $p^{\mathbf{a}}$
OP = overall power
OPF = expected fraction of correct identification
ofv = least-squares objective function value
P = general distribution
q = number of constraints
r = equations' residuals
s = number of hypothesized gross errors
t = number of nonzero eigenvalues
$\mathbf{p}^{\mathbf{r}}$ = principal component vector of vector $\mathbf{r}$
$\mathbf{p}^{\mathbf{a}}$ = principal component vector of vector $\mathbf{a}$
$\mathbf{U}_{\mathbf{r}}$ = matrix of orthonormalized eigenvectors of Φ
$\mathbf{U}_{\mathbf{a}}$ = matrix of orthonormalized eigenvectors of $\mathbf{V}$
$\mathbf{V}$ = covariance matrix of $\mathbf{a}$
$\mathbf{W}_{\mathbf{r}}$ = matrix defined by eq 5
$\mathbf{W}_{\mathbf{a}}$ = matrix defined by eq 11
$\hat{x}$ = reconciled measurements
$\mathbf{y}$ = vector of measurements

Greek Symbols

Ψ = measurement error covariance matrix
Φ = residual covariance matrix
$\Lambda_{\mathbf{r}}$ = diagonal matrix of the eigenvalues of Φ
$\Lambda_{\mathbf{a}}$ = diagonal matrix of the eigenvalues of $\mathbf{V}$
$\delta(n \times 1)$ = measurement biases
$\gamma(m \times 1)$ = leaks
ϵ = vector of random measurement errors
τ_D, τ_E = threshold tolerances

$\mu_{\mathbf{r}}$ = expected value of $\mathbf{r}$
θ_1 = vector with elements of δ and γ
τ_c = critical value for the test statistic
Σ = covariance matrix of $\hat{x}$
$\Sigma_c^{old} = \Sigma_c$ for a resolved problem
$\Sigma_c^{new} = \Sigma_c$ for a new problem
ζ = prescribed tolerance
$\lambda_{\mathbf{a},i}$ = ith eigenvalue of $\mathbf{V}$

Literature Cited

(1) Reilly, P.; Carpani, R. Application of Statistical Theory of Adjustments to Material Balances. 13th Canadian Chemical Engineering Conference, Montreal, Canada, 1963.

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

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

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

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

(6) Rollins, D.; Davis, J. Unbiased Estimation of Gross Errors in Process Measurements. *AIChE J.* **1992**, *38*, 563.

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

(8) Tong, H.; Crowe, C. Detecting Persistent Gross Errors by Sequential Analysis of Principal Components. *Comput. Chem. Eng.* **1996**, *S20*, S733.

(9) Tong, H.; Bluck, D. 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, May 1998.

(10) Keller, J.; Darouach, M.; Karzala, G. Fault Detection of Multiple Biases or Process Leaks in Linear Steady-State Systems. *Comput. Chem. Eng.* **1994**, *18*, 1001.

(11) Kim, I.; Kang, M.; Park, S.; Edgar, T. Robust Data Reconciliation and Gross Error Detection: The Modified MIMT using NLP. *Comput. Chem. Eng.* **1997**, *21*, 782.

(12) Sánchez, M.; Romagnoli, J. Monitoreo de Procesos Continuos: Análisis Comparativo de Técnicas de Identificación y Cálculo de Bias en los Sensores. AADECA 94—XIV Simposio Nacional de Control Automático, Buenos Aires, Argentina, Sept 1994.

(13) Bagajewicz, M.; Jiang, Q. Gross Error Modeling and Detection in Plant Linear Dynamic Reconciliation. *Comput. Chem. Eng.* **1998**, *22* (12), 1789—1810.

(14) Bagajewicz, M.; Jiang, Q.; Sánchez, M. Removing Singularities and Assessing Uncertainties in Two Efficient Gross Error Collective Compensation Methods. Submitted for publication in *Chem. Eng. Commun.*

(15) Sánchez, M.; Romagnoli, J.; Jiang, Q.; Bagajewicz, M. Simultaneous Estimations of Biases and Leaks in Process Plants. Submitted for publication in *Comput. Chem. Eng.*

(16) Gertler, J.; Li, W.; Huang, Y.; McAvoy, T. Isolation Enhanced Principal Component Analysis. IFAC Workshop on On-Line-Fault Detection and Supervision in the Chemical Process Industries, Solaize, Lyon, France, May 1998.

(17) Vedam, H.; Venkatasubramanian, V. Automated Interpretation of PCA-based Process Monitoring and Fault Diagnosis Using Signed Digraphs. IFAC Workshop on On-Line-Fault Detection and Supervision in the Chemical Process Industries, Solaize, Lyon, France, May 1998.

(18) Jia, F.; Martin, E.; Morris, J. Nonlinear Principal Components Analysis for Process Fault Detection. *Comput. Chem. Eng.* **1998**, *S22*, 851.

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

(20) Jiang, Q.; Bagajewicz, M. On a Strategy of Serial Identification with Collective Compensation for Multiple Gross Error Estimation in Linear Steady-State Reconciliation. Submitted for publication in *Ind. Eng. Chem. Res.*

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

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

(23) Sánchez, M.; Romagnoli, J. Use of Orthogonal Transformations in Classification/Data Reconciliation. *Comput. Chem. Eng.* **1996**, *20*, 483.

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

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

(26) Rosenberg, J.; Mah, R. S. H.; Iordache, C. Evaluation of Schemes for Detecting and Identifying Gross Errors in Process Data. *Ind. Eng. Chem. Res.* **1987**, *26*, 555—564.

Received for review July 20, 1998
Revised manuscript received February 2, 1999
Accepted February 2, 1999

IE9804677