Gross Error Detection When Variance-Covariance Matrices Are Unknown

D. K. Rollins and J. F. Davis
Dept. of Chemical Engineering, The Ohio State University, Columbus, OH 43210

Equations introduced here identify measurement biases and process leaks, when gross errors exist in measured process variables and the variance-covariance matrix of the measurements, Σ, is unknown. Σ is estimated by the sample variance, S, using process data.

For an unknown Σ, the global test statistic is the well-known Hotelling T^2 statistic. Its power function has a noncentral F-distribution. For component tests used for specific identification of measurement biases and nodal leaks, two tests are presented with Σ unknown. The first test is independent of the number of component tests, k, and is given by a statistic with an F-distribution. The second test depends on k and has a student t-distribution. The power functions for both component tests are provided. Process examples and a Monte Carlo simulation study presented demonstrate the use and performance of these statistical equations in identifying biases and process leaks.

Introduction

Gross errors in process measurements can adversely affect process performance by causing poor process control and by contributing to inadequate diagnosis. Thus, the identification of gross measurement errors and the replacement of these measured variables with accurate estimates of their true values are important for achieving optimal process performance.

A typical assumption is that measurement variance-covariance matrix, Σ, is known (Mah and Tamhane, 1982; Tamhane and Mah, 1985; Narasimhan and Mah, 1987; Narasimhan and Mah, 1988; Crowe, 1988; Rollins and Davis, 1992). However, for actual process conditions, it is more reasonable to assume that Σ is unknown since measurement errors change frequently. Changes in the magnitude of measurements and the magnitude of measurement biases are two common causes of changing error levels. Therefore, for actual process conditions we propose only that current data be used to obtain estimates of Σ. That is, we recommend that the sample variance-covariance matrix, S (calculated by current data), be used as the estimate of Σ. No researchers, to our knowledge, have statistically adapted their GED technique for an unknown Σ.

The ability of S to accurately estimate Σ increases as N increases. In addition, the statistical results improve as N increases. Thus, it is important for N to be as large as possible. However, when N is large, it is not likely that the process measurements will vary due to measurement error only. Changing process conditions will also contribute to the variability of process variables. Both sources of variability are considered by the proposed statistical procedures.

In this article, the issue of unknown variances and covariances is discussed in the context of the unbiased estimation technique (UBET) introduced in Rollins and Davis (1992). The basic goal of the UBET is to find unbiased estimates for process variables when gross errors in the measurements exist. Accomplishing this goal not only provides accurate estimates of process variables but also allows the construction of $100(1-\alpha)\%$ confidence intervals. The UBET consists of α-level global and component tests with their appropriate power functions. The global test is used to test for the existence of one or more biased measurements or process leaks. The component tests are used to identify specific biased measurements and nodes with leaks. The power functions are used to control the probability of finding leaks and biases of specified magnitudes. The extension of the UBET for unknown Σ, therefore, includes appropriate adjustments to the global and component test statistics and their power functions.

Correspondence concerning this article should be addressed to J. F. Davis.
D. K. Rollins is presently at the Depts. of Chemical Engineering and Statistics, Iowa State University, Ames, IA 50011.

We first show, in this article, the statistical models and describe the applicable variance-covariance matrices. Secondly, we give the UBET statistical procedures when the variance-covariance matrices are estimated by sample variance-covariance matrices. Thirdly, we discuss estimates and confidence intervals when variance-covariance matrices are unknown. Lastly, we present the results of a Monte Carlo simulation study that confirm the theoretical equations and show that this technique is capable of high identification performance when multiple biased measurements and process leaks are present.

Statistical Models

Measurement model

The statistical model (which is exactly the same as the one we presented in 1992, except for the addition of τ in Eq. 2) is used to describe process measurements and physical constraints (material and energy balances):

$$\bar{y} = \mu + \epsilon \tag{1}$$

such that

$$A\mu = M\gamma + \tau \tag{2}$$

where

$$\epsilon \sim N_p(\delta, \Sigma/N) \tag{3}$$

$$\tau \sim N_q(\mathbf{0}, \Sigma_\tau/N) \tag{4}$$

$$\mathrm{COV}(A\epsilon, \tau) = \mathbf{0} \tag{5}$$

$$\bar{y}^T = [\bar{y}_1, \ldots, \bar{y}_p] \tag{6}$$

$$\bar{y}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij} \tag{7}$$

$$j = 1, \ldots, p$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_{N1} & y_{N2} & \cdots & y_{Np} \end{bmatrix} \tag{8}$$

$$M = [m_1, \ldots, m_q] \tag{9}$$

As shown by Eq. 2 (assuming generation is zero), τ represents accumulation (of mass or energy). That is, τ is a term used to describe nonsteady state conditions that can be represented by Eq. 4. Notice, $E[\tau]$ is $\mathbf{0}$. When changes in process variables are within control limits (for example, no setpoint changes), this assumption should be reasonable; especially when N is large. However, if process changes are large during the period a sample is collected, this sample should not be used. In any event, the validity of this assumption ($E[\tau] = \mathbf{0}$), should be assessed whenever this model is used. In a later section we discuss the importance of τ from the viewpoint of its variability.

The sample variance for the jth measured variable is:

$$s_{jj} = \frac{1}{N-1}\sum_{i=1}^{N} (y_{ij} - \bar{y}_j)^2 \tag{10}$$

and the sample covariance between the jth and kth measured variable is:

$$s_{jk} = \frac{1}{N-1}\sum_{i=1}^{N} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k) \tag{11}$$

where the $p \times p$ sample variance-covariance matrix is:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} = (s_{jk}). \tag{12}$$

$$j, k = 1, \ldots, p$$

Transformed measurement model

The unbiased estimation technique (UBET) is developed from the following transformation of Eq. 1 (Rollins and Davis, 1992):

$$\bar{r} = A\bar{y} + A\mu + A\epsilon = M\gamma + \tau + A\epsilon \tag{13}$$

Note that

$$\mu_r = E[\bar{r}] = M\gamma + A\delta \tag{14}$$

and

$$\mathrm{VAR}(\bar{r}) = \frac{\Sigma_r}{N} = \frac{\Sigma_\tau + A\Sigma A^T}{N} \tag{15}$$

Therefore,

$$\bar{r} \sim N_q(M\gamma + A\delta, \Sigma_r/N) \tag{16}$$

Alternatively, $\bar{r}$ can be determined from the measurement data matrix, Y, and the constraint matrix, A, as follows:

$$r = YA^T = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_{N1} & y_{N2} & \cdots & y_{Np} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{bmatrix} \tag{17}$$

$$= \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1q} \\ r_{21} & r_{22} & \cdots & r_{2q} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ r_{N1} & r_{N2} & \cdots & r_{Nq} \end{bmatrix} = [r_1, \ldots, r_q] \tag{18}$$

where

$$r_j^T = [r_{1j}, \ldots, r_{Nj}] \tag{19}$$

and the components of $\bar{r}$ are:

$$\bar{r}_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \tag{20}$$

$$j = 1, \ldots, q$$

The residual constraint data matrix, Eq. 18, will also be useful later when S_r, the estimate of Σ_r, is given.

Based on this model (Eq. 13), Σ_r has two components; a component representing measurement variability, $A\Sigma A^T$, and a component representing process variability, Σ_τ (see Eq. 15). If no significant changes occur in the measured variables while the data are being taken, it is reasonable to neglect the process variability component, Σ_τ. If, on the other hand, significant process changes occur, serious errors could result if Σ_τ is neglected. [Not only does the UBET depend on Σ_r, but also the measurement test techniques (Narasimhan and Mah, 1987). Serious errors could result from neglecting Σ_τ when using any of these techniques.]

Note that the data matrix Y (Eq. 8) has N entries for each variable. The sampling time (each value of N) is the time it takes to obtain a complete row in Y. This time can be very large if any of the measured variables are determined by laboratory analysis. As previously stated, the larger the sampling time, the more likely that there will be process variations during sampling. Hence, it may not be reasonable to assume that process variability is negligible (that is, neglect Σ_τ and set $\Sigma_r = N^{-1}A\Sigma A^T$). More specifically, it may not be reasonable to estimate Σ_r as ASA^T since this calculation involves measurement variability only. Our recommendation, therefore, is to estimate Σ_r by S_r (given below) since it takes into account both process variability and measurement variability.

S_r can be determined by using the residual constraint matrix, Eq. 18, and the components of $\bar{r}$, which are given by Eq. 20. Hence, the sample variance for the jth residual constraint is:

$$s_{rjj} = \frac{1}{N-1} \sum_{i=1}^{N} (r_{ij} - \bar{r}_j)^2 \tag{21}$$

and the sample covariance between the jth and kth residual constraint is:

$$s_{rjk} = \frac{1}{N-1} \sum_{i=1}^{N} (r_{ij} - \bar{r}_j)(r_{ik} - \bar{r}_k) \tag{22}$$

$$j, k = 1, \ldots, q$$

where the $q \times q$ sample variance-covariance matrix is:

$$S_r = \begin{bmatrix} s_{r11} & s_{r12} & \cdots & s_{r1q} \\ s_{r21} & s_{r22} & \cdots & s_{r2q} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ s_{rq1} & s_{rq2} & \cdots & s_{rqq} \end{bmatrix} = (s_{rjk}) \tag{23}$$

Unbiased Estimation Technique

Global test

For the global test, the null hypothesis is H_o: $\mu_r = \mathbf{0}$ and the alternative hypothesis is H_a: $\mu_r \neq \mathbf{0}$. If δ and γ are zero, $\mu_r = \mathbf{0}$ as shown by Eq. 14. Thus, under the null hypothesis δ and γ are zero. However, if δ or γ is not zero, the alternative hypothesis, H_a: $\mu_r \neq \mathbf{0}$ will be true; assuming, of course, that error cancellation does not occur (this is a reasonable assumption. See Rollins and Davis, 1992). When Σ_r is unknown, an appropriate α-level test is to reject H_o in favor of H_a, if and only if:

$$N\bar{r}^T S_r^{-1} \bar{r} \geq \frac{(N-1)q}{N-q} F_{q,N-q,\alpha} \tag{24}$$

Note that the statistic given by Eq. 24 is the well-known Hotelling T^2 statistic (Johnson and Wichern, 1982). The power function for the test, given by Eq. 24, is:

$$\beta = \mathcal{P}\left[N\bar{r}^T S_r^{-1} \bar{r} \geq \frac{(N-1)q}{N-q} F_{q,N-q,\alpha} \,|\, \mu_r \right]$$

$$= \mathcal{P}\left[\text{noncentral } F_{q,N-q} \geq \frac{(N-1)q}{N-q} F_{q,N-q,\alpha} \,|\, \Delta^2 \right] \tag{25}$$

with

$$\Delta^2 = N\mu_r^T \Sigma_r^{-1} \mu_r \tag{26}$$

$$S_r = \begin{bmatrix} s_{r11} & s_{r12} & \cdots & s_{r1q} \\ s_{r21} & s_{r22} & \cdots & s_{r2q} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ s_{rq1} & s_{rq2} & \cdots & s_{rqq} \end{bmatrix} = (s_{rjk}) \tag{27}$$

where Δ^2 is called the noncentrality parameter, S_r is given by Eq. 23 and $F_{q,N-q,\alpha}$ is the upper (100α)th percentile of the usual F distribution with q and $(N-q)$ degrees of freedom. [See Scheffé (1959) and Bickel and Doksum (1977) for discussions on noncentral probability distributions.] S_r, the estimate of Σ_r, can be used in Eq. 26 to calculate Δ^2. However, when S_r is used, β will not be the actual power but an approximation.

As an example, consider the process network shown in Figure 1 which was taken from Narasimhan and Mah (1987) and also used by Rollins and Davis (1992). It consists of four unit operations (that is, nodes) and seven streams which are measured. Notice that two of the streams are recycle streams. There are four independent material balance constraint equations. Therefore, $q = 4$. Using the Nag (1991) software package,

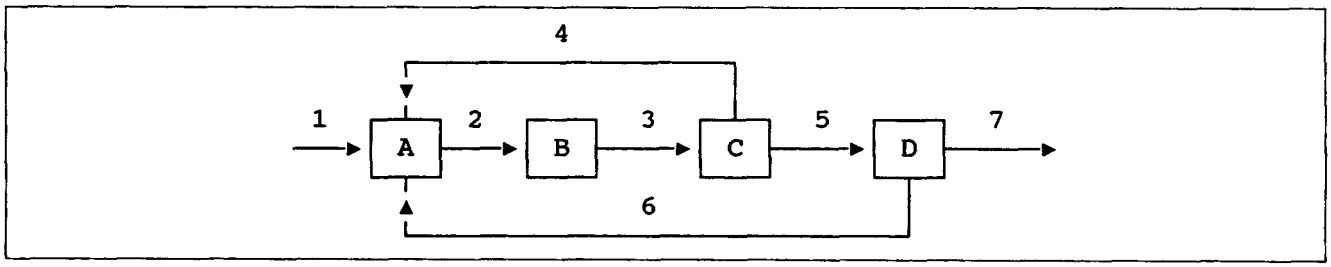

Figure 1. Recycle process network.

measured data were artificially generated six times (that is, $N=6$) for each stream and used in Eqs. 17 and 18 to give:

$$r^T=[10.5 \quad -0.46 \quad -0.13 \quad 0.55] \tag{28}$$

$$S_r=\begin{bmatrix} 2.86 & -0.91 & -0.27 & -0.39 \\ -0.91 & 1.82 & -0.61 & -1.25 \\ -0.27 & -0.62 & 0.72 & 0.46 \\ -0.39 & -1.25 & 0.46 & 1.57 \end{bmatrix} \tag{29}$$

By application of the results in Eqs. 28 and 29 the global test statistic is $N\bar{r}^T S_r^{-1}\bar{r}=1{,}427.2$. With $F_{q,N-q,\alpha}=19.25$ for $\alpha=0.05$, the critical value is $\{(N-1)q/(N-q)\}F_{q,N-q,\alpha}=192.5$. Thus, by Eq. 24, H_o is rejected and the conclusion is that at least one of the seven measured variables in Figure 1 is biased.

We now give a power calculation for the above example with $N=6$, $\delta_1=10$, $\delta_2=\delta_3=\delta_4=0$, $\Sigma=I$, and $\Sigma_r=I$ which were the conditions that generated the results given by Eqs. 28 and 29. Using the Nag (1991) algorithm for the noncentral F distribution, with $\Delta^2=N\mu_r^T\Sigma_r^{-1}\mu_r=371.4$, $\beta=0.9914$.

Component tests

Based on the statistical model given by Eq. 13, when H_o: $\mu_r=0$ is not true, at least one component of μ_r is not zero. Components of μ_r will not likely be zero if components of δ or γ are not zero (see Rollins, 1990). The basic identification mechanism of the UBET is to relate linear combinations of the components of μ_r to specific components of δ or γ. That is, to find vectors, say l_i or l_j, such that

$$l_i^T\mu_r=\delta_i \tag{30}$$

or

$$l_j^T\mu_r=\gamma_j \tag{31}$$

Thus, the component hypothesis, H_{oi}: $l_i^T\mu_r=\delta_i=0$ vs. H_{ai}: $l_i^T\mu_r=\delta_i\neq 0$ is a specific test for the presence of δ_i. Similarly, the component hypothesis, H_{oj}: $l_j^T\mu_r=\gamma_j\neq 0$ vs. H_{aj}: $l_j^T\mu_r=\gamma_j\neq 0$ is a specific test for the presence of γ_j.

Two identification approaches with values for l_i and l_j are given in Rollins and Davis (1992). The test statistics for the component tests when Σ_r is unknown are developed from the premise that $l^T\bar{r}$ is distributed $N(l^T\mu_r, N^{-1}l^T\Sigma_r l)$ for all $l^T\bar{r}$, whether Σ_r is known or unknown (Rollins and Davis, 1990; Mardia et al., 1979). The tests with Σ_r known have different distributions than the tests with Σ_r unknown due to the random nature of S_r. The power of one test depends on the number of hypotheses tested (k), but the other one does not depend on k. The test that does not depend on k will probably have the highest power when k is very large (Johnson and Wichern, 1982). The test that depends on k will be referred to as the Bonferroni test, and the test that does not depend on k will be referred to as the F test. These classifications will become apparent in a moment when the tests are introduced.

The Bonferroni test, for any l, is given as reject H_o: $l^T\mu_r=0$ in favor of H_a: $l^T\mu_r\neq 0$, if and only if:

$$\frac{\sqrt{N}|l^T\bar{r}|}{\sqrt{l^T S_r l}}\geq t_{\alpha/2k,N-1} \tag{32}$$

where $t_{\alpha/2k,N-1}$ is the upper $(100\alpha/2k)$th percentile of the t_{N-1} distribution. The statistic given in Eq. 32 is the Bonferroni statistic when Σ_r is unknown. The Bonferroni $100(1-\alpha)\%$ confidence interval for $l^T\mu_r$ is:

$$(l_i^T\bar{r}-c, \quad l_i^T\bar{r}+c) \tag{33}$$

where

$$c=t_{\alpha/2k,N-1}\sqrt{\frac{l^T S_r l}{N}} \tag{34}$$

The power function for the test given by Eq. 32 is (Scheffé, 1950; Bickel and Doksum, 1977):

$$\beta=\mathcal{P}\left[\frac{N(l^T\bar{r})^2}{l^T S_r l}\geq t^2_{\alpha/2k,N-1}\,|\,l^T\mu_r\right]$$

$$=\mathcal{P}[\text{noncentral } F_{1,N-1}\geq t^2_{\alpha/2k,N-1}\,|\,\Delta^2] \tag{35}$$

since the noncentral $t^2_{N-1}=$ noncentral $F_{1,N-1}$ and where

$$\Delta^2=\frac{N(l^T\mu_r)^2}{l^T\Sigma_r l} \tag{36}$$

For any vector l, an α-level component test derived from Eq. 24 (that is, the F test), for H_o: $l^T\mu_r=0$ vs. H_a: $l^T\mu_r\neq 0$, is to reject H_o in favor of H_a, if and only if:

$$\frac{N(l^T\bar{r})^2}{l^T S_r l}\,\frac{N-q}{q(N-1)}\geq F_{q,N-q,\alpha} \tag{37}$$

A corresponding simultaneous $100(1-\alpha)\%$ confidence interval for $l^T\mu_r$ is:

$$(l^T\bar{r}-b,\quad l^T\bar{r}+b) \tag{38}$$

where

$$b=\sqrt{\frac{q(N-1)}{N(N-q)}F_{q,N-q,\alpha}\,l^TS_rl} \tag{39}$$

The power function for this test is given as:

$$\beta=\mathcal{P}\left[\frac{N(l^T\bar{r})^2}{l^TS_rl}\geq\frac{(N-1)q}{N-q}F_{q,N-q,\alpha}\,|\,l^T\mu_r\right]$$

$$=\mathcal{P}\left[\text{noncentral } F_{1,N-1}\geq\frac{(N-1)q}{N-q}F_{q,N-q,\alpha}\,|\,\Delta^2\right] \tag{40}$$

Bonferroni intervals will be shorter than those given by Eq. 33 when c is less than b. In other words, Bonferroni intervals are smaller if $t_{\alpha/2k,N-1}$ is less than $[(q(N-1))/(N-q)(F_{q,N-q,\alpha}]^{1/2}$. For $k\leq q$ confidence statements, Bonferroni intervals are usually shorter than those given by Eq. 30 (Johnson and Wichern, 1982).

For the results and conditions given in Example 1, the Bonferroni and F component tests, and their confidence intervals, and their power results will now be given with $l^T=[1\ 0\ 0\ 0]$ (that is, a mass balance on node 1). Also, for the purpose of illustration, we are assuming that $k=5$ (for example, that the four nodal balances and the overall mass balance were tested for closure). For the F test, Eq. 37, the following quantities are determined: $l^T\bar{r}=10.5$, $l^TS_rl=2.86$, and $N(l^T\bar{r})^2/(l^TS_rl)\cdot(N-q)/\{q(N-1)\}=22.99>F_{q,N-q,\alpha}=19.25$. Thus, the F test conclusion is to reject H_o: $l^T\mu_r=\mu_A=0$; that is, it concludes that at least one of the measurements at node 1 is biased as described at the beginning of this section. In addition, with the appropriate substitutions into Eq. 38, $b=9.57$, and a 95% confidence interval for $\delta_1-\delta_2+\delta_4+\delta_6$ (that is, the δ's at this node) is (0.8898, 20.037). The power of this test is $\beta=0.99$ with $\Delta^2=150$.

Similarly, for the Bonferroni test, since $N^{1/2}|l^T\bar{r}|/(l^TS_rl)^{1/2}=6.19>t_{\alpha/10,6}=4.03$, it also rejects H_o: $l^T\mu_r=\mu_A=0$ at the $\alpha=0.05$ level. The value of c in Eq. 34 is 2.16 and thus, the 95% confidence interval is (8.8, 12.6). Note that this interval is considerably shorter than the F test interval. Finally, with $\Delta^2=150.0$, the power of this test is $\beta=1.000$ which is greater than the power of the F test, as expected.

Estimates and confidence intervals for the μ_i's

When Σ and Σ_r are unknown, estimates may be obtained for μ_i, $i=1,\ldots,p$, by using the estimators in Rollins and Davis (1992) with S replacing Σ and S_r replacing Σ_r. However, the determination of exact $100(1-\alpha)$% confidence intervals for the μ_i's is not possible because the distributions of these estimators are not known. In cases when N is large, a practical assumption is that $\Sigma=S$ and $\Sigma_r=S_r$, since they will be close to each other with high probabilities (Mardia et al., 1979). That is, the estimated values of the variance-covariance matrices may be assumed to be the true values, and confidence intervals [with levels slightly less than $100(1-\alpha)$] can be determined using the equations given in Rollins and Davis (1992).

Performance via a simulation study

In order to demonstrate the relative performance of the above component tests for identification and to verify equations in the previous section, we ran a simulation study similar to that in Rollins and Davis (1992). We used the process given by Figure 1 and ran cases with one $\delta\neq0$, with two δ's$\neq0$, and one $\gamma\neq0$. We also varied N but $\Sigma=I$, $\Sigma_r=I$, and $\alpha=0.05$, throughout.

The results of this study are presented using three measures of performance: the average type I error (AVTI), the overall power (OP), and the overall performance (OPF). Each result for all three were calculated using 10,000 cases of simulated data from the following equations (Rollins and Davis, 1992):

$$\text{AVTI}=\frac{\text{No. of zero } \delta\text{'s and } \gamma\text{'s wrongly identified}}{\text{No. of simulation trials (10,000)}} \tag{41}$$

$$\text{OP}=\frac{\text{No. of nonzero } \delta\text{'s and } \gamma\text{'s correctly identified}}{\text{No. of nonzero } \delta\text{'s and } \gamma\text{'s simulated}} \tag{42}$$

$$\text{OPF}=\frac{\text{No. of trials with perfect identification}}{\text{No. of simulation trials (10,000)}} \tag{43}$$

Note that OP and OPF values are bounded between 0 and 1. OPF results were also determined using the power functions given in the previous section and probability theory as described in Rollins and Davis (1992). Thus, OPF results of the simulation study served to check theoretically determined values which also checked the validity of the equations in the previous section. Henceforth, the theoretically determined OPF values will be denoted as OPF*.

The results of this study are given in Tables 1-6. Tables 1, 3, and 5 are identification results using the Bonferroni test and Tables 2, 4, and 5 are the ones using the F test. In Tables 1

Table 1. Bonferroni Test

i	N	OPF	AVTI	OP	OPF*
1	5	0.38	0.03	0.39	0.19
1	10	0.97	0.04	0.99	0.96
1	15	0.97	0.04	1.00	0.97
2	5	0.56	0.03	0.58	0.49
2	10	0.97	0.04	1.00	0.97
2	15	0.97	0.04	1.00	0.97
3	5	0.65	0.02	0.67	0.60
3	10	0.98	0.03	1.00	0.97
3	15	0.98	0.03	1.00	0.97
4	5	0.49	0.04	0.50	0.38
4	10	0.97	0.05	1.00	0.97
4	15	0.97	0.05	1.00	0.97
5	5	0.57	0.03	0.58	0.48
5	10	0.97	0.04	1.00	0.97
5	15	0.97	0.04	1.00	0.97
6	5	0.49	0.04	0.51	0.38
6	10	0.97	0.05	1.00	0.97
6	15	0.97	0.05	1.00	0.97
7	5	0.46	0.02	0.47	0.29
7	10	0.98	0.03	0.99	0.96
7	15	0.98	0.03	1.00	0.97

$\delta_i=5$, $\alpha=0.05$, $k=5$; all four nodal balances and the overall mass balance are used.

Table 2. F Test

i	N	OPF	AVTI	OP	OPF*
1	5	0.00	0.00	0.00	0.00
1	10	0.73	0.00	0.73	0.70
1	15	0.99	0.01	1.00	0.99
2	5	0.00	0.00	0.00	0.00
2	10	0.89	0.00	0.89	0.88
2	15	1.00	0.00	1.00	0.99
3	5	0.00	0.00	0.00	0.00
3	10	0.95	0.00	0.95	0.94
3	15	1.00	0.00	1.00	0.99
4	5	0.00	0.00	0.00	0.00
4	10	0.86	0.00	0.86	0.85
4	15	1.00	0.00	1.00	0.99
5	5	0.00	0.00	0.00	0.00
5	10	0.92	0.00	0.92	0.91
5	15	1.00	0.00	1.00	0.99
6	5	0.00	0.00	0.00	0.00
6	10	0.86	0.00	0.86	0.85
6	15	1.00	0.01	1.00	0.99
7	5	0.00	0.00	0.00	0.00
7	10	0.78	0.00	0.78	0.76
7	15	1.00	0.00	1.00	0.99

$\delta_i = 5$, $\alpha = 0.05$, $k = 5$; all four nodal balances and the overall mass balance are used.

Table 3. Bonferroni Test

i	N	OPF	AVTI	OP	OPF*
2	5	0.48	0.03	0.50	0.41
2	10	0.97	0.05	1.00	0.97
2	15	0.97	0.05	1.00	0.97
3	5	0.43	0.04	0.44	0.29
3	10	0.97	0.06	0.99	0.96
3	15	0.97	0.06	1.00	0.97

$\gamma_i = 5$, $\alpha = 0.05$, $k = 5$; all four nodal balances and the overall mass balance are used.

and 2, only one $\delta \neq 0$ for each of the seven process variables and there are three values of N (5, 10, and 15) for each value of $\delta \neq 0$. Tables 3 and 4 contain results when γ_2 or γ_3 is not zero with N also equal to 5, 10, and 15. When either one δ or γ is nonzero, as in Tables 1-4, identification is performed assuming only one δ or γ is nonzero. Hence, the hypothesis tests for the four nodal balances and the overall mass balances are sufficient for accurate identification. Thus, identification in Tables 1-4 were made testing only these hypotheses.

In Tables 5 and 6, a maximum of two δ's are assumed to be possible. We determined that nine hypothesis tests were sufficient to cover all possibilities. They were the previous five plus mass balances around nodes A, B and C; B and C; C and D; and A and B. Note also that not all possible combinations with two δ's $\neq 0$ are represented in Tables 5 and 6. Only the combinations that can give specific conclusions for the δ's are represented. For example, the case with δ_1 and $\delta_6 \neq 0$ leads to a conclusion that at least two δ's are not zero with δ_1, δ_6 and δ_7 when correct conclusions are made for all nine hypothesis tests. For more information about this identification limitation see Rollins and Davis (1992).

When only one δ (Tables 1 and 2) or one γ (Tables 3 and 4) is nonzero, Tables 1 to 4 show, for small N, that the OPF Bonferroni performance is significantly better than the F test performance. However, as N increases, the difference becomes insignificant. Therefore, since the F test is not affected by the number of hypotheses tested, these results show that for a sufficiently large N it may be possible to have high identification accuracy without concern for the size of k. In addition, the tables with one nonzero δ or γ (Tables 1 to 4) and the tables with two nonzero δ's (Tables 5 and 6) show that OPF approximates the actual OPF very well, particularly for values above 0.9. Thus, the equations for the test statistics and the

Table 4. F Test

i	N	OPF	AVTI	OP	OPF*
2	5	0.00	0.00	0.00	0.00
2	10	0.81	0.00	0.81	0.74
2	15	1.00	0.01	1.00	0.99
3	5	0.00	0.00	0.00	0.00
3	10	0.78	0.00	0.78	0.76
3	15	0.99	0.01	1.00	0.99

$\gamma_i = 5$, $\alpha = 0.05$, $k = 9$; all four nodal balances and the overall mass balance are used.

Table 5. Bonferroni Test

i	j	OPF	AVTI	OP	OPF*
1	2	0.98	0.00	1.00	0.97
1	3	0.97	0.02	0.99	0.98
1	4	0.99	0.03	1.00	0.98
1	5	0.97	0.03	0.99	0.98
2	5	0.98	0.00	1.00	0.97
2	6	0.99	0.02	1.00	0.98
2	7	0.99	0.00	1.00	0.97
3	5	0.99	0.02	1.00	0.98
3	6	0.96	0.03	0.97	0.97
3	7	0.97	0.00	0.99	0.97
4	7	0.99	0.03	1.00	0.98
5	7	0.98	0.00	1.00	0.97

$\delta_i = 15$, $\delta_j = 10$, $\alpha = 0.05$, $k = 9$, $N = 10$; the following mass balances were considered: A, B, C, D, ABCD, ABC, BC, CD, and AB.

Table 6. F Test

i	j	OPF	AVTI	OP	OPF*
1	2	0.91	0.00	0.91	0.89
1	3	0.82	0.00	0.82	0.80
1	4	1.00	0.00	1.00	0.99
1	5	0.81	0.00	0.82	0.80
2	5	0.89	0.00	0.90	0.89
2	6	0.90	0.00	0.91	0.89
2	7	1.00	0.00	1.00	0.99
3	5	0.96	0.00	0.96	0.95
3	6	0.70	0.00	0.70	0.62
3	7	0.81	0.00	0.81	0.80
4	7	1.00	0.00	1.00	0.99
5	7	0.96	0.00	0.96	0.95

$\delta_i = 15$, $\delta_j = 10$, $\alpha = 0.05$, $k = 9$, $N = 10$; the following mass balances were considered: A, B, C, D, ABCD, ABC, BC, CD, and AB.

power functions appear to be correct and can be used to approximate the overall performance accurately. Finally, all these tables also show that this approach, with unknown variances and covariances, can accurately identify biases and leaks.

Summary

This article has dealt with statistical gross error detection (GED) when variance-covariance matrices are unknown. This condition has not been adequately confronted in the chemical engineering literature even though it is very common. Tamhane and Mah (1985) cited two obstacles to gross error detection when variance-covariance matrices are unknown: (1) estimating their values from current data and (2) modifying the statistical equations. This work has addressed both obstacles. We addressed the first one by introducing S_r as an estimate for Σ_r. S_r takes into account not only the measurement variability, Σ, but also the process variability, Σ_τ. With regard to the second obstacle, this work presented statistical tests, confidence intervals and power functions when variance-covariance matrices are unknown and showed that they can be very effective in identifying biased measurements and process leaks.

Acknowledgment

We would like to acknowledge partial support for this project by The Mobil Research and Development Corporation, Princeton, NJ, and The Shell Development Company, Houston, TX. We are grateful to Melissa Rainey for developing the algorithms used in the examples and the simulation study.

Notation

a_{ij} = element of the matrix A
A = process constraint matrix
AVTI = average type I error
b = constant in Eq. 31
COV = covariance
c = constant in Eq. 36
$E[x]$ = expected value of x
e_i = vector with a 1 in the ith place and 0 elsewhere
$F_{q,N-q,\Delta}$ = noncentral $F_{q,N-q}$ random variable with noncentrality parameter Δ^2
$F_{q,N-q,\alpha}$ = upper (100α)th percentile of the $F_{q,N-q}$ distribution
H_a = alternative hypothesis
H_o = null hypothesis
k = number of tests or confidence intervals
l = vector used for making linear combinations of means or measurements
l_i = ith element of l
m_j = vector with a value in the jth place and zeros elsewhere
m_j = constant for leak in node j
M = matrix of leak constants
N = number of samples
N_p = multivariate normal p distribution
N_q = multivariate normal q distribution
OP = overall power
OPF = overall performance
$\mathcal{P}$ = probability
p = number of measurements or variables
q = number of constraints
$\bar{r}$ = transformed measurement vector ($A\bar{y}$)
r_i = ith column of YA^T
$\bar{r}_{ij}$ = element of the matrix YA^T
$\bar{r}_i$ = average residual constraint of the ith variable
s_{ij} = elements in S
s_{rij} = elements in S_r
S = sample variance-covariance measurement matrix
S_r = sample variance-covariance matrix for $\bar{r}$
$t_{N-1,\Delta}$ = noncentral t random variable with noncentrality parameter Δ^2
$t_{\alpha/2q,N-1}$ = upper 100($\alpha/2q$)th percentile of the t_{N-1} distribution
$\bar{y}$ = vector of measurement means
$\bar{y}_j$ = average measurement for the jth variable
y_{ij} = element of the Matrix Y
Y = measurement data matrix

Greek letters

α = type I error level
β = power, 1 − the type II error level
β_i = power for the ith test
γ = unknown vector of process leak constants
δ = vector of measurement biases
δ_i = ith element of δ
Δ^2 = noncentrality parameter
ϵ = vector of random errors
μ = unknown vector of true means
μ_r = $E[\bar{r}]$
Σ = variance-covariance measurement matrix
Σ_r = variance-covariance matrix for $\bar{r}$
Σ_τ = variance-covariance matrix for τ
τ = unknown vector for the accumulation of mass or energy
$\chi^2_{q,\alpha}$ = upper (100α)th percentile of the χ^2_q distribution

Superscripts

T = "transpose"

Special symbol

$\sim$ = "is distributed"

Literature Cited

Bickel, P. J., and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-day, San Francisco (1977).

Crowe, C. M., "Recursive Identification of Gross Errors in Linear Data Reconciliation," *AIChE J.*, **34**(4), 541 (1988).

Heenan, W. A., and R. W. Serth, "Detecting Errors in Process Data," *Chem. Eng.* (Nov., 1986).

Johnson, R. A., and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, NJ (1982).

Mah, R. S. H., "Data Screening," *Proc. FOCAPO Meeting* (1987).

Mah, R. S. H., and A. C. Tamhane, "Detection of Gross Errors in Process Data," *AIChE J.*, **28**(5), 828 (1982).

Mardia, K. V., J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, New York (1979).

NAG, The Numerical Algorithms Groups, Inc., Downers Grove, IL (1987).

Narasimhan, S., and R. S. H. Mah, "Generalized Likelihood Ratio Methods for Gross Error Identification," *AIChE J.*, **33**(9), 1514 (1987).

Narasimhan, S., and R. S. H. Mah, "Generalized Likelihood Ratios for Gross Error Identification in Dynamic Processes," *AIChE J.*, **34**(8), 1321 (1988).

Rollins, D. K., "Unbiased Estimates of Measured Process Variables When Measurement Biases and Process Leaks Are Present," PhD Diss., The Ohio State Univ., Columbus (1990).

Rollins, D. K., and J. F. Davis, "Unbiased Estimation of Gross Errors in Process Measurements," *AIChE J.*, **38**(4), 563 (1992).

Scheffé, H., *The Analysis of Variance*, Wiley, New York (1959).

Tamhane, A. C., and R. S. H. Mah, "Data Reconciliation and Gross Error Detection in Chemical Process Networks," *Technometrics*, **27**(4), 409 (1985).

Manuscript received Feb. 13, 1992, and revision received Jan. 20, 1993.