Recursive Identification of Gross Errors in Linear Data Reconciliation

In the reconciliation of measurements of flows and concentrations so that they conform to conservation laws and other constraints, any gross errors in the measurements must be identified in order that they can be either corrected or deleted.

A new method is derived for the recursive prediction of the changes in the objective function, and of the statistical tests for the measurements, which would result from the deletion of suspect measurements. Inverses of large matrices are not required and the reconciliation can also be easily calculated for any set of deletions. It is shown that the decrease in the objective function caused by deletion of a single measurement equals the square of the corresponding maximum power measurement statistic, calculated prior to that deletion. An algorithm for the detection of suspect sets of gross errors, whose deletion leads to acceptable values of all statistical tests and process flow rates, is proposed and illustrated.

Cameron M. Crowe
Department of Chemical Engineering
McMaster University
Hamilton, Ontario, Canada L8S 4L7

Introduction

The computation of the steady state material and energy balances from measurements of the flows and concentrations of streams in a process is an essential step in the monitoring of process performance. Unfortunately, the plant measurements are corrupted by fluctuations in the process, by instrument or analytical errors, by unknown leaks, and by departures from steady state. The errors may be random or biased, so the latter must be identified and corrected, or the measurements discarded, before any reconciliation of the data based on random variation may be done.

Reconciliation of process measurements consists in adjusting those data, in some sense minimally, so that the adjusted values obey the conservation laws and any other constraints imposed on the process. Examples of such constraints are the nonnegativity of concentrations and the sum of all component flows being equal to the total flow in a stream.

If one or more gross errors is present in the measurements, the adjustments made to all of the measurements will be strongly affected and the statistical assumption of purely random error is not valid. Thus one wishes to detect the presence of gross errors and to identify the offending measurements. Various statistical tests have been proposed to detect gross errors, namely the global chi-square test and the species imbalance test (Reilly and Carpani, 1963), and the measurement test (Mah and Tamhane, 1982; Crowe et al., 1983). Several algorithms for detecting gross errors have recently been published by Rosenberg et al. (1987), Iordache et al. (1985), and Serth and Heenan (1986). These authors have tested and compared algorithms using simulated measurements with random error, to which selected gross errors were added. Romagnoli and Stephanopoulos (1981) suggested reconciling the data by starting with only one balance and progressively adding balances until the statistical tests were violated. It is not clear whether the sequence in which the balances are added would alter the detectibility of gross errors.

The effect of deleting a set of measurements has been evaluated by repeating the reconciliation after deletion and assessing the changes in the various statistical tests. An efficient method of obtaining these subsequent reconciliations from the original one has been described by Ripps (1965). In this method, the well-known formula for the inverse of a bordered matrix (Lapidus, 1962) is used to calculate the inverse of a matrix, obtained by removing a set of bordering rows and columns from a larger matrix whose inverse is already known.

It is the purpose of this paper to present a method of predicting the effect of deleting any given set of measurements on the statistical tests without carrying out the full reconciliations. The aim is to find one or more sets of deleted measurements that satisfy all of the statistical tests. The derivation presented also leads directly to the calculation of the measurement statistics and of the reconciliation itself.

Linear Reconciliation Problem

The theory will be restricted here to the linear problem, that is, where the measurements consist of species and total flow rates. For simplicity of notation, we will assume that the balances contain only measured flow rates. This is not a limitation since Crowe et al. (1983) showed how the unmeasured flow rates could be eliminated from the balance equations. The reconciliation problem can then be defined as

$$\min_{a} J = a^T \Sigma^{-1} a$$

subject to

$$B(\tilde{x} + a) = 0 \tag{1}$$

Here, $\tilde{x}$ is an n-vector of measurements, a is the corresponding vector of adjustments, and B is an $m \times n$ matrix, of rank m, containing the incidence matrix for the balances on species, together possibly with rows for additional equality constraints. It is assumed that the measurements have a known or estimated positive definite $n \times n$ variance matrix Σ. The solution, following Crowe et al. (1983), can be written as

$$a = -\Sigma B^T H^{-1} e \tag{2}$$

with imbalance error vector

$$e = B\tilde{x} \tag{3}$$

and positive definite matrix

$$H = B\Sigma B^T \tag{4}$$

From the formula for the variance of a linear combination of jointly distributed random variables (Scheffe, 1959, p. 8), if the expectation of e is zero and the variance of the measurements is Σ, then the variance of e is H. One can verify by substitution of Eq. 2 that the value of the objective function at the solution is

$$J = e^T H^{-1} e \tag{5}$$

Furthermore, from Scheffe (p. 418, problem V.2), it has a chi-square distribution. This is then the basis for the global chi-square test and also for the premise that if all measurements in gross error have been correctly deleted, the objective function after deletion will be below the threshold for the chi square with the appropriate confidence level and number of degrees of freedom. It does not follow, if the chi-square test is satisfied, that there are no gross errors since one gross error may exist among a large set of random measurements. We thus prefer to use further specific tests to diagnose the measurements and the species imbalances individually.

Deletion of a Set of Measurements

Suppose that we wish to assess the effect of deleting a particular set of ℓ measurements on the objective function. We partition matrix B as

$$B = [B' \mid B''] \tag{6}$$

where the ℓ columns of B'' correspond to the deleted values. We assume that the rank of B'' is ℓ ($<m$). Since these deleted values are effectively unmeasured, following Crowe et al. (1983), we define matrix R^T, $(m - \ell) \times m$, by

$$R^T B'' = 0 \tag{7}$$

with R of rank $(m - \ell)$. It is shown below that if the columns of B'' were linearly dependent, deletion only of measurements corresponding to a maximal set of independent columns from B'' would give the same value of the objective function as the deletion of all of them. Then the solution to the case with deletion is obtained from Eqs. (1–4) by replacing B by $R^T B$ and thus e and H in Eqs. 3 and 4 by

$$e_d = R^T e \tag{8}$$

and

$$H_d = R^T H R \tag{9}$$

respectively. Then the change in the objective function J is

$$\begin{aligned} \Delta J &= e_d^T H_d^{-1} e_d - e^T H^{-1} e \\ &= -e^T M e \end{aligned} \tag{10}$$

with

$$M = [H^{-1} - R H_d^{-1} R^T] \tag{11}$$

We wish to find an expression for M in terms of quantities already available. First, we note from Eqs. 7 and 9 that

$$MHR = 0 \tag{12}$$

and

$$HMB'' = B'' \tag{13}$$

From Eq. 12, the null space of M has dimension at least $(m - \ell)$, the rank of HR. Hence, the rank of M is at most ℓ. From Eq. 13, HM has the columns of B'' as independent eigenvectors with as many eigenvalues equal to unity as ℓ, the rank of B''. Thus the rank of HM, and hence that of M, is exactly ℓ. It is readily verified, as shown in the Appendix, that Eqs. 12 and 13 are uniquely satisfied by

$$M = H^{-1} B'' G_\ell^{-1} B''^T H^{-1} \tag{14}$$

where the matrix G_ℓ is defined by

$$G_\ell = B''^T H^{-1} B'' \tag{15}$$

The reduction in the objective function can then be written from Eqs. 10 and 14 as

$$\Delta J = -(e^T H^{-1} B'') G_\ell^{-1} (B''^T H^{-1} e) \tag{16}$$

Now if we were to add to B'' an additional measurement to be deleted whose column b in B' is dependent on those of B'', the matrix R would not change since it would already satisfy

$\boldsymbol{R}^T\boldsymbol{b} = \boldsymbol{0}$. Thus, from Eqs. 8 and 9, neither $\boldsymbol{e}_d$ nor $\boldsymbol{H}_d$ would change, so that the addition of a dependent column to $\boldsymbol{B}''$ leaves $\boldsymbol{M}$, and hence the change in objective function, unaltered.

In particular, when a single measurement is deleted, Eq. 16 simplifies to

$$\delta J = -(\boldsymbol{b}^T\boldsymbol{H}^{-1}\boldsymbol{e})^2/(\boldsymbol{b}^T\boldsymbol{H}^{-1}\boldsymbol{b}) \tag{17}$$

where $\boldsymbol{b}$ is the single column of $\boldsymbol{B}$ corresponding to the deleted measurement.

Aside from vector-matrix multiplications, the only computational effort needed is the inversion of $\boldsymbol{G}_\ell$. This can be done recursively from the inverse of $\boldsymbol{G}_{\ell-1}$ with the formula for the inverse of a bordered matrix (Lapidus, 1962). However, since the number of measurements to be deleted at one time will not likely be large, the savings in computer time and programming effort will be small.

Equation 16 then allows one to predict the reduction in the objective function as a result of each of a sequence of trial deletions, singly, two at a time, and so on. Any set of values whose deletion does not reduce the objective function enough to satisfy the chi-square test cannot contain all of the gross errors. Among the sets of ℓ deletions, those that do satisfy the chi-square test are examined further to establish whether each such set also satisfies all of the measurement and imbalance tests.

As a final check, for each set that so far satisfies all criteria, all flows including the unmeasured ones are computed. The results are checked for reasonableness, especially to weed out cases that give negative flows. Any set which, when deleted, leads to negative flows indicates that there is a serious flaw in the remaining measurements and that it is unlikely that all of the correct gross errors have been deleted, given reasonable variances. Negative flows could arise if the variances assigned to those flows, especially small ones, are too large. If no sets of the size ℓ can be found which satisfy all statistical tests and which do not lead to negative flows, then more measurements should be deleted if possible.

Imbalance Test

Since the variance of the imbalance vector $\boldsymbol{e}$ is $\boldsymbol{H}$, the variance of any linear combination $\boldsymbol{w}^T\boldsymbol{e}$ is $\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{w}$. Thus a general imbalance test for such a linear combination is given by

$$z_e(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{e}/(\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{w})^{1/2} \tag{18}$$

If it is assumed that $\boldsymbol{e}$ is normally distributed with expected value of zero, then $z_e(\boldsymbol{w})$ has zero mean and unit variance. Usually, $\boldsymbol{w}$ would be a unit vector but this is not necessary.

After the deletion of ℓ measurements, the imbalance statistic can be written for any linear combination $\boldsymbol{v}$ as

$$\begin{aligned} z_{ed}(\boldsymbol{v}) &= \boldsymbol{v}^T\boldsymbol{e}_d/(\boldsymbol{v}^T\boldsymbol{H}_d\boldsymbol{v})^{1/2} \\ &= \boldsymbol{v}^T\boldsymbol{R}^T\boldsymbol{e}/(\boldsymbol{v}^T\boldsymbol{R}^T\boldsymbol{H}\boldsymbol{R}\boldsymbol{v})^{1/2} \end{aligned} \tag{19}$$

This is very similar to Eq. 18 except that $\boldsymbol{w}$ is replaced by $\boldsymbol{R}\boldsymbol{v}$ with $\boldsymbol{v}$ typically a unit vector. Thus the application of the imbalance test to the deleted case is straightforward, without having to perform the reconciliation, although the matrix $\boldsymbol{R}$ would be required.

Matrix $\boldsymbol{R}$, for the additional deletion of a measurement corresponding to column $\boldsymbol{b}$ of $\boldsymbol{B}'$, can be formed readily from a previous $\boldsymbol{R}_1$ prior to that deletion ($\boldsymbol{I}$ for no deletion at all) from

$$\boldsymbol{R}^T = [\boldsymbol{c}_1 \mid -\gamma\boldsymbol{I}]\boldsymbol{R}_1^T \tag{20}$$

where

$$\boldsymbol{c} = [\gamma \mid \boldsymbol{c}_1^T]^T = \boldsymbol{R}_1^T\boldsymbol{b} \tag{21}$$

with $\gamma \neq 0$. If the first element of $\boldsymbol{c}$ is zero, the columns of the term in square brackets in Eq. 20 would be permuted so that $\boldsymbol{c}_1$ is placed in the position occupied by an arbitrary choice of $\gamma \neq 0$ in $\boldsymbol{c}$.

Measurement Test

In a similar manner, and under the assumption that the imbalance is normally distributed with mean zero and variance $\boldsymbol{H}$, the adjustment vector $\boldsymbol{a}$ was shown by Mah and Tamhane (1982) and by Crowe et al. (1983) to have mean zero and a singular variance matrix

$$\boldsymbol{Q} = \boldsymbol{\Sigma}\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B}\boldsymbol{\Sigma} \tag{22}$$

so that elements of $\boldsymbol{a}$ could be tested against the unit normal variate by

$$z_{a,j} = \boldsymbol{u}_j^T\boldsymbol{a}/(\boldsymbol{u}_j^T\boldsymbol{Q}\boldsymbol{u}_j)^{1/2} \tag{23}$$

where $\boldsymbol{u}_j$ is the jth unit vector, that is, the jth column of the identity matrix.

Mah and Tamhane (1982) showed that a test of maximum power (MP) for the measurements before any deletions, could be defined by

$$\begin{aligned} z^*_{a,j} &= \boldsymbol{u}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}/(\boldsymbol{u}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_j)^{1/2} \\ &= -\boldsymbol{b}_j^T\boldsymbol{H}^{-1}\boldsymbol{e}/(\boldsymbol{b}_j^T\boldsymbol{H}^{-1}\boldsymbol{b}_j)^{1/2} \end{aligned} \tag{24}$$

from Eqs. 2 and 22 and with $\boldsymbol{b}_j = \boldsymbol{B}\boldsymbol{u}_j$, the particular column of $\boldsymbol{B}$. Tamhane (1982) proved that if a single measurement were in gross error, its MP statistic would exceed or equal in expected absolute value the MP statistic of any other measurement. They further showed that the MP statistic for the measurement in gross error would exceed or equal in absolute expected value any other unit normal measurement statistic with an arbitrary square nonsingular matrix replacing $\boldsymbol{\Sigma}^{-1}$. It should be noted that Almasy and Sztano (1975) defined this same statistic and observed its property of maximality for the value having gross error. After deletion, the equivalent statistic for a remaining undeleted measurement would be

$$z^*_{ad,j} = -\boldsymbol{b}_j^T\boldsymbol{R}\boldsymbol{H}_d^{-1}\boldsymbol{R}^T\boldsymbol{e}/(\boldsymbol{b}_j^T\boldsymbol{R}\boldsymbol{H}_d^{-1}\boldsymbol{R}^T\boldsymbol{b}_j)^{1/2} \tag{25}$$

The change in the numerator between Eqs. 24 and 25 is

$$\begin{aligned} \Delta(\text{num}) &= \boldsymbol{b}_j^T[\boldsymbol{H}^{-1} - \boldsymbol{R}\boldsymbol{H}_d^{-1}\boldsymbol{R}^T]\boldsymbol{e} \\ &= \boldsymbol{b}_j^T\boldsymbol{M}\boldsymbol{e} \end{aligned} \tag{26}$$

The change in the square of the denominator is similarly

$$\Delta(\text{denom}^2) = -\boldsymbol{b}_j^T\boldsymbol{M}\boldsymbol{b}_j \tag{27}$$

Thus the measurement test for any of the undeleted values can be calculated quickly using Eq. 14 for M. Note that the statistic, Eq. 25, would be undefined for any of the deleted values because of Eqs. 6 and 7.

We note that the prediction of the change in any measurement test does not require the calculation of the matrix R, whereas the prediction of the imbalance test does. Some savings in computational effort can then be achieved by only calculating the imbalance test if all other tests have been satisfied for a particular set of deleted values.

Simple Calculations of ΔJ from z^*_{ad} for an Additional Deletion

We now can see that given a reconciliation with a particular set of measurements, the reduction in the objective function caused by the deletion of any one measurement is precisely equal to the square of the corresponding maximum power measurement statistic prior to its deletion. Thus, for any measurement in a given reconciliation, comparing the MP measurement statistics from Eq. 24 with the change in the objective function resulting from the deletion of that measurement from Eq. 17, with $b = b_j$, we find that

$$\delta J = -(z^*_{a,j})^2 \tag{28}$$

Then given the reconciliation, we know from the measurement statistics which single deletions (if any) will lead to a sufficiently large reduction in the objective function to bring it below the tabulated chi-square value.

It follows that after predicting the effect of any particular set of deleted measurements on the measurement statistics of the remaining ones, the additional reduction in the objective function by one more deletion is similarly the square of the corresponding measurement statistic prior to its deletion.

An alternative equation for the change in the objective function upon deletion of a given set of measurements corresponding to B'' can be obtained from Eqs. 16 and 24 as

$$\Delta J = -Z''^T \Gamma^{-1} Z'' \tag{29}$$

where

$$Z'' = [z^*_{a,1} z^*_{a,2}, \ldots, z^*_{a,\ell}]^T \tag{30}$$

the vector of measurement statistics corresponding to columns of B'', before deletion, for those values to be deleted and Γ is the matrix whose elements are

$$\Gamma_{ij} = b_i^T H^{-1} b_j / [(b_i^T H^{-1} b_i)(b_j^T H^{-1} b_j)]^{1/2} \tag{31}$$

with b_i, b_j as columns of B'' for $i, j = 1, 2, \ldots, \ell$.

Calcuation of an Additional Deletion from a Previously Deleted Set

Let us assume that the objective function and the measurement statistics have been calculated for the deletion of ℓ measurements corresponding to the columns of B''. We wish to obtain a simple formula for calculating the measurement statistic of each remaining measurement, after the further deletion of the value corresponding to column b_q of B'.

To express the change in the numerator and the square of the denominator of the measurement statistic for the further deletion of b_q, we note that the form of the equations is preserved but with the replacement of e by e_d, Eq. 8, and of H by H_d, Eq. 9. The inverse of H, in the equations for the measurement statistics and the adjustments, is replaced by

$$N \equiv R H_d^{-1} R^T \tag{32}$$

as seen in Eq. 25 compared to Eq. 24. Then the derivation to obtain Eq. 14 for M can be followed analogously to give, with the measurements from B'' already deleted,

$$M_n = N b_q b_q^T N / (b_q^T N b_q) \tag{33}$$

for the additional deletion of the value corresponding to b_q.

The change in the numerator of $z^*_{ad,j}$ is then, by analogy to Eq. 26,

$$\delta(\text{num}_j) = b_j^T M_n e \tag{34}$$

which, with Eq. 33, becomes

$$\delta(\text{num}_j) = (b_j^T N b_q)(b_q^T N e)/(b_q^T N b_q) \tag{35}$$

Similarly, from Eqs. 27 and 33,

$$\delta(\text{denom}_j^2) = -(b_j^T N b_q)^2/(b_q^T N b_q) \tag{36}$$

Then, from Eqs. 25 and 32,

$$z^*_{ad,j} = -(b_j^T N e)/(b_j^T N b_j)^{1/2} \tag{37}$$

prior to deletion of b_q, so that the new value of the statistic is

$$z^*_{an,j} = (z^*_{ad,j} - \Gamma_{n,jq} z^*_{ad,q})/(1 - \Gamma^2_{n,jq})^{1/2} \tag{38}$$

with

$$\Gamma_{n,jq} = (b_j^T N b_q)/[(b_j^T N b_j)(b_q^T N b_q)]^{1/2} \tag{39}$$

with M and H^{-1} already calculated, the matrix N also available from Eqs. 11 and 32 as

$$N = H^{-1} - M \tag{40}$$

so that the new measurement statistics can easily be obtained. We also note from Eqs. 2 and 11 that the change in the adjustment of the jth remaining measurement after deleting the measurements from the columns of B'' is

$$\Delta a_j = u_j^T \Sigma B^T M e \tag{41}$$

so that the additional change in a_j caused by the further deletion of the value represented by column b_q is

$$\delta a_j = u_j^T \Sigma B^T M_n e \tag{42}$$

$$= u_j^T \Sigma B^T N b_q (b_q^T N e)/(b_q^T N b_q)$$
$$= -u_j^T \Sigma B^T N b_q z^*_{ad,q}/(b_q^T N b_q)^{1/2} \tag{43}$$

We have already seen in Eq. 28 that the change in the objec-

tive function due to an additional single deletion is the square of the corresponding measurement statistic prior to that deletion. Thus, the additional change in the objective function caused by the further deletion of the measurement with column b_q is

$$\begin{aligned}\delta J &= -e^T M_n e \\ &= -(b_q^T N e)^2/(b_q^T N b_q) \\ &= -(z^*_{ad,q})^2 \end{aligned} \tag{44}$$

from Eq. 37.

Dependence among Columns of B and Equality of Measurement Statistics

After the deletion of measurements corresponding to columns of B'', a measurement statistic for an undeleted value has exactly zero numerator and denominator, and is thus indeterminate, if and only if the corresponding column of B' is linearly dependent on the columns of B''. Thus, from Eq. 25 and since the matrix H_d is positive definite, the denominator is zero if and only if

$$R^T b_j = 0 \tag{45}$$

This means that the column b_j of B is a linear combination of the columns of B'', from Eq. 7 for the definition of R. Clearly, the numerator of Eq. 25 is zero for random values of e (i.e., with probability one) if and only if the denominator is zero.

As Iordache et al. (1985) showed in their theorem 3, the MP statistics of two remaining undeleted measurements are equal in absolute value if and only if the corresponding columns of $R^T B'$ are proportional to each other, that is

$$R^T(b_i - \beta b_j) = 0 \quad \text{for} \quad \beta \neq 0 \tag{46}$$

Almasy and Sztano (1975) previously also observed the sufficiency of Eq. 46 for the equality of two MP statistics. This implies that for some vector $w \neq 0$, again from Eq. 7 for the definition of R,

$$(b_i - \beta b_j) = B'' w \tag{47}$$

Such a case of equality implies that the additional deletion of either value will cause the same reduction in the objective function and the same values of the other measurement statistics since b_i would be dependent on the columns of B'' augmented by b_j, and vice versa. Furthermore, the undeleted value will have an indeterminate measurement statistic.

Then if one such value is a member of a suspect set, so should the other one be, and no discrimination between them is possible unless one or both lead to unacceptable values of the reconciled or unmeasured flows. The implication of this is that one needs to assess the effect of deleting only one of a set of measurements with equal values of their statistics and thus that B'' should always have full column rank.

Algorithm for Gross Error Detection

There are two elements of a strategy for an algorithm for the identification of gross errors. The first is to predict the reduction in the objective function and to find the minimum number of deletions required in order that there be at least one set which reduces values of J to less than the chosen criterion level of the chi-square test. Then any such set of deleted values is rejected if any undeleted value has a measurement statistic that is greater than a chosen level of the unit normal variate. Each set that still remains is then subjected to the imbalance test and rejected if it fails. Finally, for each remaining set the flow rates are estimated and assesses for consistency with any other criteria, such as nonnegativity.

The second element of the strategy uses the values of the measurement statistics to flag absolute equality and thus to reduce the number of combinations of deletions that need to be examined, as well as to avoid dealing with singular matrices. After the deletion of any specific set of measurements, the measurement statistics of the remaining values again predict the effect of one further deletion.

There are alternatives for calculating the effect of a trial deletion of a set of ℓ measurements. The first uses the base case itself to find the objective function from Eq. 16 and the measurement statistics from Eqs. 24, 26, and 27. This would require the storage of the common matrix-vector products involving B, H^{-1}, and e but would not need the storage of the measurement statistics for all deletions of $(\ell - 1)$ values.

The second alternative would economize on the computational effort by using the previously stored values of the measurement statistics for each of the sets of $(\ell - 1)$ deletions. The objective function and the statistics for the additional deletion of each remaining measurement would be obtained from Eqs. 44 and 38, respectively. The amount of storage needed for a large process with many levels of deletion could become difficult to manage.

Before conducting a series of trial deletions, one may ask whether there is a preferred order in which to do this. Whereas the final results would be independent of the order in which the trials were done, there are differences in the number of trials needed. Specifically, there are two general approaches, breadth-first and depth-first. In addition, one can examine the more likely sources of gross error before the less likely by listing the species flow rates in descending order of absolute value of their MP measurement statistics.

In the breadth-first approach, the trial deletions are carried out at each level before moving to the next one. Thus, all single deletions are evaluated before calculating those for deletion of sets of two, and so on. This approach has the advantage of needing only to delete down to the level of the minimum number of deletions required to give a satisfactory reconciliation. However, if the above alternative is used, the results at each level must be stored for use at the next level. In all cases, the sets to be deleted are chosen so that no case is treated twice.

In the depth-first approach, each species is taken in turn and the sequence of deletions involving it singly, in pairs, and so on is evaluated before the next species is examined. This tree of deletions for a species is ended when a set is found whose deletion gives a satisfactory reconciliation among the remaining measurements or when the maximum number of deletions has been made. In this approach, we must examine more cases than in the breadth-first approach but less storage is needed and the results can immediately be used to calculate the next level with one additional deletion. In the algorithm that follows, we have chosen to use the breadth-first approach.

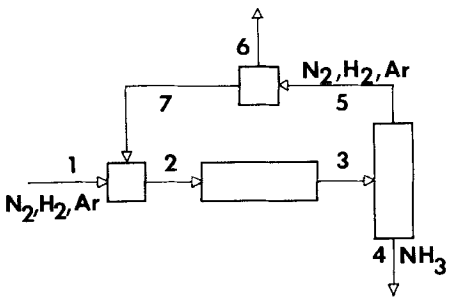

Figure 1. Ammonia synthesis loop flow sheet.

The algorithm is set out as follows:

Algorithm A

1. (a) Enter input data: B, Σ, and $\tilde{x}$.

(b) Compute H and its inverse.

(c) Compute a and e from Eqs. 2, 3, and $\hat{x} = \tilde{x} + a$, the reconciled flows.

(d) Compute the objective function J from Eq. 5 and the MP measurement statistics from Eq. 24.

(e) IF $J < \chi^2$ for m DF (degrees of freedom) at chosen confidence level AND all measurement and imbalance statistics (Eqs. 18 and 24) are less than the criterion level of the unit normal variate, AND no flows are unacceptable, THEN there are apparently no gross errors present. Go to 10.

2. (a) Set ℓ_{max} ($\leq m - 1$), the maximum number of deletions.

(b) Set $\ell = 0$.

3. (a) Order the measured species flow rates such that their absolute measurement statistics are monotonic decreasing. (This will place flows that are more likely to be in gross error near the head of the list.)

4. (a) Set p = number of the initial measurement statistics that differ in absolute value by more than a small relative tolerance, say 1.E-10 in double precision computations. Flag those subsets whose absolute values differ by less than the tolerance and use only one of them for deletion trials.

5. $\ell = \ell + 1$

6. DO, for $k = 1$ to p:

(a) Select the kth distinct set, S_k, of ℓ measurements, taking account of the flagged subsets.

(b) Compute the change ΔJ in objective function if S_k were deleted, from Eq. 44 for a particular species j and from the previous change with the deletion of all species in Sk except species j.

(c) Compute the measurement test values for all remaining unflagged measurements and for allowable representatives of flagged variables, from Eq. 38.

(d) IF $(J + \Delta J) <$ tabulated $\chi^2_{\alpha,(m-\ell)}$ AND
IF all measurement tests are passed THEN
Compute imbalance test values
IF all imbalance tests are passed THEN
Mark S_k as suspect
END IF
END IF

(e) For each subset of undeleted measurements with absolute values of the MP statistic within the tolerance, flag the union of that subset and the deleted variables in S_k, if it is not already flagged. From this flagged subset, no choice of $(\ell + 2)$ values should be deleted and only one choice of $(\ell + 1)$ values needs to be deleted together.

END DO

7. For each S_k marked as suspect, compute the reconciled flow rates from the changes in adjustments, using Eq. 41 or 43 with that set deleted. Verify whether any adjusted measurements are unacceptable. If so, remove the mark on S_k since this set does not lead to an acceptable reconciliation.

8. IF no sets S_k are still marked, AND $\ell < \ell_{max}$,
Set p = number of distinct subsets of $(\ell + 1)$ variables, taking into account all flagged subsets. Go to 5.
END IF

9. For each marked set S_k in which there is a measurement whose statistic was flagged as differing in absolute value from that of another measurement by less than the tolerance, generate additional sets by replacing each flagged variable by such a previously disallowed choice. For each such set, compute the unmeasured flows and discard any set producing unacceptable values since it does not give an acceptable reconciliation.

10. Output results and stop.

It is important to note that we cannot rely upon identifying suspect gross errors one at a time if there are two or more gross errors. It is shown in the example below that such serial deletion would fail to find the correct set of gross errors. There is no guarantee that there will be only one suspect set of measurements so that if more than one set is identified, further measurement data may need to be examined to establish whether such sets continue to be suspect. It is doubtful that any algorithm could be devised which made no errors of misidentification, either type I or II. In

Table 1. Data for Ammonia Synthesis Loop Example

Base Case: +20% gross error in $NH_3^{(4)}$, +10% in $N_2^{(1)}$

	$N_2^{(1)}$	$H_2^{(1)}$	$Ar^{(1)}$	$N_2^{(2)}$	$Ar^{(2)}$	$N_2^{(3)}$	$NH_3^{(4)}$	$H_2^{(5)}$
B	0	0	0	1	0	−1	−0.5	0
	1	0	0	−1	0	0.98	0	0
	0	1	0	0	0	0	−1.5	−0.02
	0	0	1	0	−0.02	0	0	0
"True"								
x	33.0	99.0	0.4	105.1	20.08	73.57	63.06	220.7
$\tilde{x}$	36.69	101.54	0.4	104.76	20.87	76.10	76.04	212.86
$\hat{x}$	36.92	110.57	0.42	109.18	21.00	73.74	70.89	211.15
z_a^*	−2.08	4.15	0.94	2.20	−0.94	−2.04	−3.73	−4.15
			(z_c = 2.63 for s' = 6, = 2.49 for s' = 4)					

J = 29.30 (cf. chi-square = 9.49 with 4 DF at 95% confidence)

Table 2. Values of Objective Function J with Trial Deletion

Measurements Deleted	Degrees of Freedom, DF	Change in J	J	X^2 0.05
None	4	—	29.30	9.49
$H_2^{(1)}$	3	17.23	12.07	7.81
$NH_3^{(4)}$		13.90	15.40	
$N_2^{(2)}$		4.82	24.48	
$N_2^{(1)}$		4.34	24.96	
$N_2^{(3)}$		4.18	25.12	
$H_2^{(1)}$, $NH_3^{(4)}$	2	22.29	7.01	5.99
$H_2^{(1)}$, $N_2^{(2)}$		23.88	5.42#	
$H_2^{(1)}$, $N_2^{(1)}$		18.70	10.52	
$H_2^{(1)}$, $N_2^{(3)}$		24.46	4.84#	
$NH_3^{(4)}$, $N_2^{(2)}$		18.14	11.16	
$NH_3^{(4)}$, $N_2^{(1)}$		25.67	3.63#	
$NH_3^{(4)}$, $N_2^{(3)}$		17.19	12.11	
$N_2^{(2)}$, $N_2^{(1)}$		8.10	21.20	
$N_2^{(2)}$, $N_2^{(3)}$		8.10	21.20	
$N_2^{(1)}$, $N_2^{(3)}$		8.10	21.20	

#Objective function < chi-square criterion.

the end, the measuring procedures and instruments will have to be reexamined in order to determine which set of measurements is responsible for the gross errors.

Application to an Example: Ammonia Synthesis Loop

As an example of the application of the algorithm, let us consider the ammonia synthesis loop used by Crowe et al. (1983). The flow sheet is shown in Figure 1. The "measurement" data were generated from values that obey the balances by the addition of normally distributed random noise with the same variance structure as in Crowe et al. Then gross errors were added to selected values. The "true" and the perturbed data used are given in Table 1 for a particular pair of gross errors, as well as the reconciled flow rates and the measurement statistics. Clearly the value of the objective function J exceeds the chi-square for 4 DF at 95% confidence. We see as expected from the proportionality of the respective columns in matrix $\boldsymbol{B}$ that the deletion of either hydrogen value will lead to the same results, as will the deletion of either argon value.

Selected results of the algorithm applied to this example are given in Tables 2 and 5. It is seen in Table 2 that no single deletion reduces the objective function enough but that three different deleted pairs do lead to a sufficient reduction and are thus marked as suspect. It is also notable that the same reduction in the objective function results from the deletion of any pair of the three nitrogen flows. This is directly linked to fact that the deletion of any one nitrogen flow leads to equality of the other two nitrogen MP statistics, as seen in Table 3.

In testing the measurement statistics, at an overall confidence level of 95%, we use the formula of Mah and Tamhane (1982) to relate the confidence level of an individual statistic to the overall level, namely

$$\alpha^* = 1 - (1 - \alpha)^{1/s'} \tag{48}$$

Here, α is the probability of wrongly rejecting the null hypothesis that there is no gross error, and is set to 0.05, while α^* is the corresponding probability level for each statistic. The integer s' ($\leq n$), according to Iordache et al. (1985), respresents the number of measurement statistics with distinct absolute values. They

Table 3. Measurement Statistics for Trial Deletions

Single Deletion: Sets $\{H_2^{(1)}, H_2^{(5)}\}$, $\{Ar^{(1)}, Ar^{(2)}\}$ flagged

Value	$N_2^{(1)}$	$H_2^{(1)}$	$Ar^{(1)}$	$N_2^{(2)}$	$Ar^{(2)}$	$N_2^{(3)}$	$NH_3^{(4)}$	$H_2^{(5)}$
				Delete $N_2^{(1)}$				
Z_{ad}^*	—	3.80	0.63	1.94	**	−1.94	−4.62	**
				Delete $H_2^{(1)}$				
Z_{ad}^*	1.25	—	1.40	2.58	**	−2.69	−2.25	**
				Delete $N_2^{(2)}$				
Z_{ad}^*	−1.81	4.37	0.17	—	**	1.81	−3.65	**
				Delete $N_2^{(3)}$				
Z_{ad}^*	−1.98	4.51	0.24	1.98	**	—	−3.61	**
				Delete $NH_3^{(4)}$				
Z_{ad}^*	−3.43	2.90	1.86	2.06	**	−1.82	—	**

Deletion of Sets of Two: Set $\{N_2^{(1)}, N_2^{(2)}, N_2^{(3)}\}$ flagged

Value	$N_2^{(1)}$	$H_2^{(1)}$	$Ar^{(1)}$	$N_2^{(2)}$	$Ar^{(2)}$	$N_2^{(3)}$	$NH_3^{(4)}$	$H_2^{(5)}$
				Delete $N_2^{(1)}$, $NH_3^{(4)}$ (Suspect I)				
Z_{ad}^*	—	−1.57	1.58	1.57	1.58	−1.57	—	1.57
Z_{ad}	—	0.54	1.39	0.93	−0.73	−1.57	—	1.57
				Delete $H_2^{(1)}$, $N_2^{(3)}$ (Suspect II)				
Z_{ad}^*	1.91	—	0.52	−1.91	−0.52	—	−1.91	—
Z_{ad}	2.19	—	1.52	−0.79	−0.59	—	−1.91	—
				Delete $H_2^{(1)}$, $N_2^{(2)}$ (Suspect III)				
Z_{ad}^*	2.05	—	0.53	—	−0.53	−2.05	−2.05	—
Z_{ad}	2.32	—	1.57	—	−0.53	−2.05	−2.05	—
				Delete $H_2^{(1)}$, $NH_3^{(4)}$ (Not suspect)				
Z_{ad}^*	−2.41	—	1.91	2.41	−1.91	−2.41	—	—
Z_{ad}	0.16	—	1.55	1.69	−0.57	−2.41	—	—

—Indeterminate because deleted.
**Not necessary to calculate since previously flagged.

Table 4. Imbalance Statistics for Suspect Deleted Pairs ($z_c = 2.235$)

R^TB								
$N_2^{(1)}$	$H_2^{(1)}$	$Ar^{(1)}$	$N_2^{(2)}$	$Ar^{(2)}$	$N_2^{(3)}$	$NH_3^{(4)}$	$H_2^{(5)}$	z_{ed}
				Suspect I: $\{N_2^{(1)}, NH_3^{(4)}\}$				
0	−1	0	3	0	−3	0	0.02	−1.06
0	0	1	0	−0.02	0	0	0	−1.08
				Suspect II: $\{H_2^{(1)}, N_2^{(3)}\}$ or $\{H_2^{(5)}, N_2^{(3)}\}$				
1	0	0	−0.02	0	0	−0.49	0	−2.13
0	0	1	0	−0.02	0	0	0	−1.08
				Suspect III: $\{H_2^{(1)}, N_2^{(2)}\}$ or $\{H_2^{(5)}, N_2^{(2)}\}$				
1	0	0	0	0	−0.02	−0.5	0	−2.26
0	0	1	0	−0.02	0	0	0	−1.08

also pointed out that the rank of the variance of z_a^*, which is equal to the number of degrees of freedom, has not been shown to provide a less conservative value for s'. The criteria z_c for a unit normal variance are then

s'	1	2	3	4	5	6
α^*	0.05	0.025	0.017	0.013	0.010	0.0085
z_c	1.96	2.235	2.39	2.49	2.57	2.63

at the overall 95% confidence level. Then a measurement statistic is too large to be random if it exceeds z_c in absolute value.

It can be seen in Tables 1 and 3 that if the MP measurement statistic exceeds z_c for s' equal to the degrees of freedom, it also exceeds z_c for the number of distinct absolute MP statistics. For deletion of two measurements, $s' = 2$ in any event. In Table 3, of the three suspect pairs, none has an MP measurement statistic that is larger than z_c, although the value of z_{ad} for $N_2^{(1)}$ does exceed z_c. If the imbalance statistics are examined in Table 4, we might discard $\{H_2^{(1)}, N_2^{(2)}\}$ for having a slightly too large imbalance statistic.

Table 5. Reconciled and Estimated Flow Rates

Stream	N_2	H_2	Ar	NH_3
		Suspect I: $\{N_2^{(1)}, NH_3^{(4)}\}$		
1	34.08	102.04	0.41	0
2	106.89	310.75	20.57	0
3	74.30	212.96	20.57	65.19
6	1.49	4.26	0.41	0
7	72.81	208.70	20.15	0
		Suspect IIa: $\{H_2^{(1)}, N_2^{(3)}\}$		
1	38.16	114.72	0.41	0
2	103.95	323.33	20.62	0
3	67.13	212.86	20.62	73.64
6	1.34	4.26	0.41	0
7	65.78	208.60	20.21	0
		Suspect IIb: $\{H_2^{(5)}, N_2^{(3)}\}$		
1	38.16	103.83	0.41	0
2	103.95	−221.33	20.62	0
3	67.13	−331.84	20.62	73.64
6	1.34	−6.64	0.41	0
7	65.78	−325.16	20.21	0
		Suspect IIIa: $\{H_2^{(1)}, N_2^{(2)}\}$		
1	38.24	114.43	0.41	0
2	112.70	323.03	20.65	0
3	75.97	212.86	20.65	73.45
6	1.52	4.26	0.41	0
7	74.45	208.60	20.24	0
		Suspect IIIb: $\{H_2^{(5)}, N_2^{(2)}\}$		
1	38.24	103.94	0.41	0
2	112.70	−201.44	20.65	0
3	75.97	−311.61	20.65	73.45
6	1.52	−6.23	0.41	0
7	74.45	−305.38	20.24	0

If the three suspect pairs are ranked in ascending order of their values of J, one sees that their maximum absolute values of both the measurement and imbalance statistics are also in ascending order. The prime suspect would then be $\{N_2^{(1)}, NH_3^{(4)}\}$ in having the lowest statistics. This is in fact the correct choice since gross errors of +10 and +20% were applied respectively to these values.

The reconciliations of the three suspect pairs lead to no inconsistent results. However, one sees in Table 5 that the two additional suspects, with $H_2^{(1)}$ replaced by $H_2^{(5)}$, do give negative values for the unmeasured hydrogen flows. One further observation is that the MP measurement statistics uniformly exceed the corresponding statistic, Eq. 23, for the prime suspect but not for the other pairs. Whether this occurs more generally in other processes remains to be established.

Finally, we note that we would have been misled in identifying the correct gross errors if we had accepted $H_2^{(1)}$, the value causing the greatest reduction in J among all single deletions, as being one of the measurements in gross error. The wrong suspect pairs would then have been found if $H_2^{(1)}$ were always included.

Conclusions

Formulas have been developed to predict the effect of deleting any set of measurements on the objective function and on the measurement test statistics. These formulas can be used without having to compute the reconciliation for each case of deletion. In particular, equations are presented with which one can calculate the effect on the objective function, the measurement statistics, and the adjustments of the additional deletion of a further measurement.

An algorithm is proposed for identifying sets of suspect measurements which appear to cause the violation of the statistical tests, in that the deletion of such a set removes that violation. The algorithm not only applies the new formulas but also limits the number of trial deletions required by taking account of equal absolute measurement statistics. In the ammonia example, only six out of eight single deletions and 13 out of 28 pairs needed to be examined. The correct pair of measurements in gross error was identified as the most likely suspect. The performance of this algorithm needs to be tested and compared to other pub-

lished algorithms, using actual or simulated data, in order to assess its ability correctly to identify gross errors and to avoid errors of type I (wrongly finding a truly random error to be a gross error) and type II (wrongly finding a truly gross error to be random).

Acknowledgment

The author wishes to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada through an Operating Grant. He is grateful to A.N. Hrymak for helpful comments and to W. Holly for carrying out the computations.

Notation

a = vector of adjustments to flow measurements, $n \times 1$
B = matrix in balance Eq. 1, $m \times n$
B' = columns of B not deleted, $m \times n - \ell$
B'' = columns of B deleted, $m \times \ell$
b = single column of B
e = vector of imbalances, Eq. 3, $m \times 1$
G = matrix defined in Eq. 15, $\ell \times \ell$
$H = B\Sigma B^T$, $m \times m$
I = identity Matrix
J = objective function, Eq. 1
M = matrix defined in Eq. 11, $m \times m$
N = matrix defined in Eq. 32, $m \times m$
Q = singular variance matrix of a, Eq. 22, $n \times n$
R = matrix with columns orthogonal to B'', Eq. 7, $m \times m - \ell$
v = arbitrary vector, Eq. 19, $m - \ell \times 1$
w = arbitrary vector, Eq. 18, $m \times 1$
x = vector of flow measurements, $n \times 1$
z = unit normal variate, scalar, or vector

Greek letters

β = nonzero scalar, Eq. 46
γ = nonzero scalar, Eq. 20
Δ = change from deletion of ℓ values corresponding to the columns of B''
δ = change resulting from additional single deletion
Γ = matrix defined in Eq. 31, $\ell \times \ell$
Σ = variance of $\tilde{x}$, $n \times n$

Subscripts

a = adjustment statistic
d = after deletion of measurements corresponding to columns of B''
e = imbalance statistic
j = counter for conserved species
n = after deletion of single additional measurement
q = additional species flow rate deleted

Superscripts

(n) = stream number n
T = transpose of a matrix or vector
$\tilde{}$ = measured flow rate
$\hat{}$ = reconciled flow rate
$*$ = maximum power measurement statistic, Eq. 24

Appendix

Lemma. The matrix M that satisfies Eqs. 12 and 13 is given uniquely by Eq. 14.

Proof. That Eq. 14 for M satisfies Eqs. 12 and 13 is verified by substitution. Thus in Eq. 12

$$\mathbf{MHR} = \mathbf{H}^{-1}\mathbf{B}''\mathbf{G}_\ell^{-1}\mathbf{B}''^T\mathbf{R} = \mathbf{0} \tag{A1}$$

from Eq. 7. Substitution into Eq. 13 gives

$$\mathbf{HMB}'' = \mathbf{B}''\mathbf{G}_\ell^{-1}\mathbf{B}''^T\mathbf{H}^{-1}\mathbf{B}'' = \mathbf{B}'' \tag{A2}$$

from Eq. 15. The uniqueness is established by assuming that there is another symmetric matrix L ($\neq M$) that also satisfies Eqs. 12 and 13. Then from Eq. A1,

$$(\mathbf{M} - \mathbf{L})\mathbf{HR} = \mathbf{0} \tag{A3}$$

and from Eq. A2,

$$\mathbf{H}(\mathbf{M} - \mathbf{L})\mathbf{B}'' = \mathbf{0} \tag{A4}$$

From Eqs. A3 and 7 and the symmetry of M, L, and H,

$$\mathbf{H}(\mathbf{M} - \mathbf{L}) = \mathbf{B}''\mathbf{W} \tag{A5}$$

for some $m \times m$ matrix W of rank $r \leq \ell$. Then again from symmetry, this leads to

$$(\mathbf{M} - \mathbf{L}) = \mathbf{W}^T\mathbf{B}''^T\mathbf{H}^{-1} \tag{A6}$$

which in turn gives, in Eq. A4,

$$\mathbf{HW}^T\mathbf{B}''^T\mathbf{H}^{-1}\mathbf{B}'' = \mathbf{0} \tag{A7}$$

or from Eq. 15,

$$\mathbf{HW}^T\mathbf{G}_\ell = \mathbf{0} \tag{A8}$$

Since $\mathbf{G}_\ell$ and $\mathbf{H}$ are nonsingular,

$$\mathbf{W} = \mathbf{0} \tag{A9}$$

and hence from Eq. A5

$$\mathbf{M} = \mathbf{L}$$

in contradiction to the premise. This completes the proof of the lemma.

Literature Cited

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

Crowe, C. M., Y. A. Garcia Campos, and A. Hrymak, "Reconciliation of Process Flow Rates by Matrix Projection. I: The Liner Case," *AIChE J.*, **29**, 881 (1983).

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

Lapidus, L., *Digital Computation for Chemical Engineers*, McGraw-Hill, New York, 251 (1962).

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

Reilly, P. M., and R. E. Carpani, "Application of Statistical Theory of Adjustment to Material Balances," *13th Can. Chem. Eng. Conf.*, Montreal (1963).

Ripps, D. L., "Adjustment of Experimental Data," *Chem. Eng. Prog. Symp. Ser.*, **61**(55), 8 (1965).

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

Rosenberg, J., R. S. H. Mah, and C. Iordache, "Evaluation of Schemes for Detecting and Identifying Gross Errors in Process Data," *Ind. Eng. Chem. Res.*, **26**, 555 (1987).
Scheffe, H., *The Analysis of Variance*, Wiley, New York (1959).
Serth, R. W., and W. A. Heenan, "Gross Error Detection and Data Reconciliation in Steam-Metering Systems," *AIChE J.*, **32**, 733 (1986).
Tamhane, A. C., "A Note on the Use of Residuals for Detecting an Outlier in Linear Regression," *Biometrika*, **69**, 488 (1982).
Manuscript received Sept. 28, 1987, and revision received Dec. 16, 1987.