II. Numerical data used in the simulation of the distillation column (Example V)

Filter gain used to obtain Fig. 11:
$K^T = (-1.354 \quad -0.646 \quad -5.558)$
for $V_1 = \text{diag}(0, 0, 3.8)$, $V_2 = .123$
Filter gain used to obtain Fig. 12:
$K^T = (-0.081 \quad -1.614 \quad -5.020)$
for $V_1 = \text{diag}(0, 0, 3.1)$, $V_2 = .123$
Filter gain used to obtain Fig. 13:
$K^T = (-1.107 \quad -0.520 \quad -4.032)$
for $V_1 = \text{diag}(0, 0, 3.8)$, $V_2 = .123$
$K^T = (-0.044 \quad -0.023)$
for $V_1 = \text{diag}(0.001, 0.001)$, $V_2 = .123$
Filter gains used to obtain Fig. 14:
$K^T = (-4.393 \quad -2.338)$
for $V_1 = \text{diag}(0.1, 0.1)$, $V_2 = .123$

Manuscript received February 6, 1979; revision received August 24, and accepted August 27, 1979.

Statistical Analysis of Constrained Data Sets

JAY C. KNEPPER
Rio Blanco Oil Shale Co.
Denver, CO 80231
and
JOHN W. GORMAN
Amoco Oil Co.
Research and Development
Naperville, IL 60540

Practical, comprehensive computer-based methods for analyzing data sets are developed. Methods for calculating data adjustment, parameter values, variance-covariance of adjusted data and parameters and for detecting aberrant data are presented. A simple calculation algorithm and application of the methods for design of experimental measurements are proposed.

SCOPE

Least squares analysis, commonly used to fit regression equations to sets of experimental data, also provides powerful techniques for analyzing the measured data themselves, when these data can be interrelated through constraining physical laws. Least squares parameter estimation problems are typically "parameter—rich" in the sense that the data set is constrained by a single regression equation. (In this article, the term "parameter" is used in the engineering sense of a quantity to be estimated from other data, rather than in the statistical sense, wherein it applies to all estimated quantities.)

In another frequently met situation of particular interest in this work, a large set of data is interrelated by physical laws such as heat balances, material balances or kinetic equations. This situation may be thought of as being "constraint-rich", and is typical of many laboratory experiments, pilot unit tests, and commercial unit performance tests. Constraints, together with estimates of the variances of the measured data, can be used to adjust data to more accurate values and to draw conclusions about their credibility.

The data estimation techniques discussed here greatly facilitate analysis of constrained data sets by providing maximum likelihood estimates of the measured data and any parameters, by assessing the probability that there are extraordinary errors, and by providing error information about the calculated quantities for use in subsequent analysis. Further, the techniques can be used to develop experiments producing data with improved accuracy. The method can be thought of as an extension of the averaging process to situations, in which each measured quantity may enter into one or a number of physical constraints, and by which alternative values can be inferred. It can be applied, if necessary, to detect gross measurement errors and to isolate any such by using data redundancies, much as a skilled analyst would. And, replicates or near replicates are not required to judge credibility of the data values.

The net effect can be significantly increased efficiency of experimentation, offering the happy choice of either obtaining more accurate data for a given cost, or of achieving the same accuracy in the final results with less experimental cost. Because the techniques are general, they can handle data sets described as a typical linear or nonlinear regression problem, a parameter-free data adjustment problem, or any combination of these.

The basic data correction algorithm was first derived for and applied to surveying problems (Demming 1946). Data adjustment techniques were applied in the chemical engineering literature by Kuhn and Davidson (1961). They handled nonlinear problems, but did not address either estimates of the error in the adjusted data, or criteria for assessing consistency of the measured data set.

Gross errors in measured data, if undetected, tend to be spread out among the calculated adjustments to the measured data. Ripps (1967) considered the possibility of gross measurement error and suggested an algorithm for computing data adjustments when gross errors are suspected. Nogita (1972) added an univariate statistical criterion for judging the probability that a blunder exists in the data set. The specific case of simultaneous chemical reactions has been discussed, with emphasis on computational aspects, by Murthy (1973, 1974).

Madron et al. (1977) also address simultaneous chemical reactions. Of particular note in their work is the multidimensional chi-square test used to test for consistency of measured values, and the calculation of the variance-covariance matrices for all estimated data. A chi-square test can, however, be applied to the general data adjustment problem, both in detecting gross measurement error and in locating likely aberrant measurements. Further, unnecessary restriction is placed on the applicability of the test in their work. In good treatment of the general estimation problem, Britt and Leucke (1973) calculated the error structure of estimated parameters and placed special emphasis on parameter estimation in nonlinear problems.

This study attempted to develop practical and comprehensive computer based tools for analyzing data sets and planning of experimental measurements. Essential elements include the ability to handle general linear and nonlinear cases, estimates of the variance of and covariance between all estimated data, generally applicable methods for detecting inconsistency in the measured data set and for isolating any aberrant measurements, and efficient methods of computation.

Address correspondence to Knepper.

0001-1541-80-3237-0260-$00.75.

CONCLUSIONS AND SIGNIFICANCE

In the present work, the data adjustment and parameter estimation equations are developed from the theory of generalized inverses. Within this framework the data adjustment equations can be developed simply while showing clearly the roles of parameters and measured data in the solution of the problem. The convergence of nonlinear problems is examined, and an efficient computational algorithm is suggested.

Variance-covariance matrices for the calculated parameters, the adjusted data, and covariance between the two are developed. These matrices are shown to be useful in deciding the value of additional experimental measurements toward increasing accuracy. In addition, a multivariate chi-square test is described which allows a check for discrepancies in the constraints that are inconsistent with the variance-covariance matrices of the measured data. Calculation of the chi-square test statistic is independent of the class of problem and removes an unnecessary restriction on whether certain estimated quantities are measured or are estimated only from other measured values and the set of constraints (Madron et al. 1977).

Other tests, including the powerful chi-square test, are developed for spotting data items which may be grossly in error. Simple numerical examples are presented to illustrate the use of the techniques.

DERIVATION OF THE ALGORITHM

The techniques are based on adjusting a set of measured data to satisfy a set of constraints, which will usually be physical laws interrelating the data. That is, we wish to develop a set of adjusted values, $\hat{Z}$, which satisfy the vector equation $F(\hat{Z}) = 0$, where F is an m-vector and $\hat{Z}$ is an n-vector of adjusted measured values. It may also be necessary to estimate a set of parameters from the data set. In this case, the preceding set of vector equations becomes $F(\hat{Z},\hat{\theta}) = 0$, where θ is a p-vector of parameters. The sizes of the vectors Z, F and θ are related by $n \geq m > p \geq 0$.

Initially the constraints will not be satisfied exactly due to error in the measured data values, and we have

$$F(Z_0, \theta_0) = F_0 \tag{1}$$

where the subscript 0 identifies measured data values in the case of Z, initial estimates of the parameter values, and the resulting discrepancies, F_0. Since the constraint equations may be nonlinear, we expand Eq. (1) about Z_0 and θ_0 and project the result to $F(Z_1, \theta_1) = 0$. Anticipating the need for an iterative solution for nonlinear problems, we replace the subscripts 0 and 1 in the linear approximation by the iteration counters $i+1$ and i to obtain

$$F_z(Z_{i+1} - Z_i) + F_\theta(\theta_{i+1} - \theta_i) = -F_i \tag{2}$$

Here F_z and F_θ denote the $m \cdot n$ and $m \cdot p$ matrices resulting from differentiating F with respect to Z and θ, and evaluating the results at Z_i and θ_i.

Since we are interested in minimizing the amount of correction to the measured values, $Z_{i+1} - Z_0$, Eq. (2) is rewritten in the form

$$F_z(Z_{i+1} - Z_0) + F_\theta(\theta_{i+1} - \theta_i) = -F_i + F_z(Z_i - Z_0) \tag{3}$$

Note that the right-hand side is a linear approximation to $-F(Z_0, \theta_i)$. Now the data adjustment, $Z_{i+1} - Z_0$, and the discrepancies, F_i are normalized, using the variance-covariance matrix of the measured data, $R = \{\sigma_{ij}\}_{n,n}$, and the variance of the discrepancies due to measurement error, $F_z R F_Z{}^T$.* It is assumed that R is known, although in practice the elements of R will most often be estimated by sample variances. We now define $d_{i+1} = D(Z_{i+1} - Z_0)$ and $f_i = HF_i$, where $D^T D = R^{-1}$ and $H^T H = (F_z R F_z^T)^{-1}$. D and H are upper triangular matrices (Lapidus 1962). Substituting for Z and F using these definitions and multiplying the equation through by H gives

$$HF_z d_{i+1} + HF_\theta(\theta_{i+1} - \theta_i) = -f_i + HF_z D^{-1} d_i \tag{4}$$

Initially, the right-hand side of Eq. (4) is equal to $-f_0 = HF(Z_0, \theta_0)$, and is distributed approximately as $N(0,I)$. The quantity $d_i^T d_i$ is recognized as the scalar $(Z_{i+1} - Z_0)^T R^{-1} (Z_{i+1} - Z_0)$, which reduces to

$$\sum_{k=1}^{n} \frac{\{Z_i^{(k)} - Z_0^{(k)}\}^2}{\{\sigma^{(k)}\}^2}$$

when there is no covariance among the measured data values.

We seek to solve Eq. (4) according to the following strategy: (1) Choose the parameters so as to minimize the euclidian norm $\| - f_i + HF_z D^{-1} d_i - HF_\theta(\theta_{i+1} - \theta_i)\|_E$, and then (2) calculate d_{i+1} such that $\|d_{i+1}\|_E$ is minimized. More simply stated, this is: Calculate the parameters in such a way that a least squares fit of f_0 is obtained, and then compute the least squares adjustment of the measured data as weighted by the inverse of the variance in the data.

The theory of generalized inverses provides an interesting route to the solution of the problem (Pearson 1974). Consider a system of m equations in n unknowns $AX = K$, with A an $m \cdot n$ matrix of coefficients, X an n vector to be calculated, and K a constant m-vector. If a solution exists, it is given by $X = A^+K$ $(I - A^+A) + C$, where C is arbitrary. Here A^+, an $n \cdot m$ matrix, is termed the generalized inverse of A. Let A be of rank r. If $m=r$, then $A^+ = A^T(AA^T)^{-1}$; while if $n=r$, then $A^+ = (A^TA)^{-1}A^T$. If a solution exists ($n \geq m$), $X = A^+K$ is the solution of least euclidian norm. If a solution does not exist ($n < m$), then $X = A^+K$ is the least squares approximation to K.

We see that the second situation fits with the strategy for evaluating the parameters, while the first corresponds to the plan for evaluating the adjustments to the measured data. We have for the parameters: $\theta_{i+1} - \theta_i = (HF_\theta)^+ (-f_i + HF_z D^{-1} d_i)$. After constructing the indicated generalized inverse, defining $Q = (F_z R F_z^T)^{-1}$ and expressing the result in terms of Z and F we find

$$\theta_{i+1} - \theta_i = (F_\theta^T Q F_\theta)^{-1} F_\theta^T Q \{-F_i + F_z(Z_i - Z_0)\} \tag{5}$$

Inserting this result in Eq. (4), solving for d_{i+1} by means of the generalized inverse of $HF_z D^{-1}$, and returning to the variables Z and F, we obtain

$$Z_{i+1} - Z_0 = RF_Z^T Q[I - F_\theta(F_\theta^T Q F_\theta)^{-1} F_\theta^T Q] \{-F_i + F_Z(Z_i - Z_0)\} \tag{6}$$

Eqs. (5) and (6) have been derived by many investigators using the method of Lagrange multipliers. They are exact for linear problems, providing the maximum likelihood estimates for Z and θ, while satisfying all constraints without iteration. In nonlinear problems, an iterative procedure with periodic reevaluation of F_z and F_θ is required, and the resulting estimates of Z and θ will be biased. Once the calculations have converged, Z_{i+1} becomes $\hat{Z}$, and θ_{i+1} is identified as $\hat{\theta}$. Necessary conditions for the existence of the inverses appearing in Eqs. (5) and (6) have been discussed (Britt and Leucke 1973).

* By the law of propagation of error (Bennett and Franklin 1954, Davies 1957)

For $(HF_\theta)^+$ and $(HF_zD^{-1})^+$ to exist as defined above, F_θ must be of rank p, and F_z of rank m. In complex problems it is often difficult to specify an independent set of constraints. Twenty constraints, 30 measured variables, and 10 parameters can easily be involved in balances around chemical pilot plants or commercial process units. Hence, at the expense of additional computation, computer runs can be saved by the fact that the generalized inverse of an $n \cdot m$ matrix A of rank r, where $r < m$ and $r < n$, is given by $A^+ = C^+(B^+AC^+)^{-1}B^+$ (Pearson 1974). Here B is any $m \cdot r$ matrix of independent rows and C is an $r \cdot n$ matrix of independent columns. C^+ and B^+ can be calculated as described above. To implement this procedure, a routine to calculate the rank of F_z and F_θ and identify the independent columns and rows is needed.

Through the use of the theory of generalized inverses, the workings of the estimation Eqs. (5) and (6) are brought clearly to light. That is, the parameters are first estimated by a least squares technique. The estimated parameter values are independent of the estimates of the adjustment to the measured data to the extent that $F_i - F_z(Z_i - Z_0)$ approximates $F(Z_0, \theta_i)$. Then, once the least squares fit of the parameters is obtained, the measured values are adjusted such that the weighted sum of the adjustment is minimized. An improperly specified variance-covariance matrix of the measured values or an improper model used as a constraint can lead to inappropriate adjustment of affected data values.

COMPUTATIONAL ASPECTS

Problems of practical interest can easily reach a size such that the solution of Eqs. (5) and (6) requires a significant amount of computing time on even the largest machines. Computing time per iteration will increase, as a rough guide, as the cube of the number of variables. Here the nature of the iterative process is explored with the aim of developing an efficient algorithm for computing nonlinear problems.

Defining $E' = RF_z^TQ[I - F_\theta B']$, $B' = (F_\theta^TQF_\theta)^{-1}F_\theta^TQ$, $E = E'F_z$, and $B = B'F_z$, Eqs. (5) and (6) become

$$\theta_{i+1} - \theta_i = -B'F_i + B(Z_i - Z_0) \tag{7}$$

and

$$Z_{i+1} - Z_0 = -E'F_i + E(Z_i - Z_0) \tag{8}$$

Eqs. (7) and (8) can be solved in a straightforward manner by using p of the constraints to calculate θ_0 and then evaluating the estimation matrices B', B, E', and E always using the latest estimates of Z and θ, and iterating until the estimated data no longer change. Most of the computation is involved in recomputing B', B, E', and E each time new vectors Z and θ are calculated. Another approach is to converge to an approximate solution using values of these matrices developed from measured data. Then the procedure, which we will refer to as the constant direction approach, is repeated, after re-evaluating the estimation matrices. Calculations can be continued in this manner until no further change is found in the estimated data.

Application of the constant direction algorithm can be considerably simplified through the use of the identities

$$EE = E,\ EE' = E',\ BE = BE' = 0,\ E'F_\theta B' = E'F_\theta B = 0$$

which can be verified by direct computation. Applying Eq. (8) for k steps gives

$$Z_{k+1} - Z_0 = -E' \sum_{j=0}^{k} F_j \tag{9}$$

or

$$Z_{k+1} - Z_k = -E'F_k \tag{10}$$

where we have used the identities $EE = E$ and $EE' = E$. We also find, using the fact that $BE = BE' = 0$,

$$\theta_{k+1} - \theta_k = -B'F_k \tag{11}$$

In comparison with Eqs. (7) and (8), Eqs. (9) or (10) and (11) are simpler to program, require less storage, and usually will require less than half of the multiplications per iteration.

Suppose that the constraints are satisfied after k trials. At this point, the estimation matrices can be re-evaluated. Renumbering point k as point 1, Eqs. (7) and (8) can be used to evaluate Z_2 and θ_2 with $F_1 = 0$. For subsequent iteration it can be shown, using Eqs. (7) and (8) repeatedly for k steps and applying the above identities, that Eqs. (9) or (10) and (11) can be used for subsequent iterations, i.e. with $k > 1$, to converge to new $\hat{Z}$ and $\hat{\theta}$ vectors. Each time new estimation matrices are evaluated, the direction of the search is refined such that d^Td is minimized. A check to see if d^Td is no longer decreasing significantly as $\hat{Z}$ and $\hat{\theta}$ are refined is a useful criterion for convergence.

The constant direction algorithm can thus be summarized as follows:

1. Evaluate θ_0 by some technique such as rough guesses or by finding a θ_0 vector which satisfies p of constraints exactly. Set $Z_1 = Z_0$, $\theta_1 = \theta_0$ and $k = 1$.
2. Evaluate successive estimates of Z and θ using Equations (10) and (11) until the sequence converges to $F(\hat{Z},\hat{\theta}) = 0$.
3. Re-evaluate E, E', B, and B' using $\hat{Z}$ and $\hat{\theta}$ from step 2.
4. Check to see if d^Td has been reduced significantly. If no, calculations are complete; if yes, proceed with step 5.
5. Set $\hat{Z} = Z_1$, $\hat{\theta} = \theta_1$, and $i = 1$ and evaluate Z_2 and θ_2 using Eqs. (7) and (8).
6. Set $k = 2$ and return to step 2.

Because of the effort required to evaluate E, E', B, and B' each time new estimates of Z and θ are produced, it is likely that the constant direction algorithm, implied by Eqs. (9) or (10) and (11), will usually be more efficient than straightforward computation based on Eqs. (7) and (8). It is, however, impossible to state that this or any single algorithm is always most efficient is solving the general class of nonlinear problems. Further, the degree of stability of neither algorithm has been addressed. The nature of changes in the vector F_i from iteration to iteration can be observed and, if stability is a problem, the adjustments $Z_{i+1} - Z_i$ and $\theta_{i+1} - \theta_i$ can be modulated by a factor between 0 and 1. Other, more sophisticated techniques for dealing with instabilities have been described (Himmelblau 1970).

ERROR STRUCTURE OF THE ESTIMATED QUANTITIES

An important motivation for applying data analysis and correction techniques is the availability of a more accurate set of measurements and parameters which satisfy the constraint equations. The error structure of the estimated quantities is important for subsequent use of the data. Further, hypothetical runs simulating data obtained from a given experiment can be used to decide if a worthwhile increase in accuracy would result from either making certain measurements more accurately, or through measuring additional pieces of data so that additional constraints can be imposed on the data set. This application complements the results of Draper and Hunter (1966) through which experiments are designed in sensitive regions for effective estimation of parameters.

To derive the necessary relations, following Britt and Leucke (1973), we expand the constraints about the true solution vectors for the measured data and parameters, $\tilde{Z}$ and $\tilde{\theta}$, respectively. We find, analogous to Eq. (3),

$$F_z(\hat{Z} - Z_0) + F_\theta(\hat{\theta} - \theta_0) = F_z(\tilde{Z} - Z_0) \tag{12}$$

and we have from Eq. (7),

$$\hat{\theta} - \tilde{\theta} = -B'F_z(Z_0 - \tilde{Z}) \tag{13}$$

and from Eq. (8), $\hat{Z} - Z_0 = E(\tilde{Z} - Z_0)$
or

$$\hat{Z} - \tilde{Z} = (I - E)(Z_0 - \tilde{Z}) \tag{14}$$

Hence the variance-covariance matrix for the estimated parameters is calculated from the expected value of $(\hat{\theta} - \tilde{\theta})(\hat{\theta} - \tilde{\theta})^T$ as

$$Ev\ \{(\hat{\theta}-\tilde{\theta})\,(\hat{\theta}-\tilde{\theta})^T\} = B'F_zRF_z^TB'^T = (F_\theta^TQF_\theta)^{-1} \quad (15)$$

as obtained by Britt and Leuke (1973).

By the same method, results for other variance-covariance matrices of interest are

$$Ev\ \{(\hat{Z}-\tilde{Z})\,(\hat{Z}-\tilde{Z})^T\} = R - RF_z^TQF_zR + RF_z^TQF_\theta(F_\theta^TQF_\theta)^{-1}F_\theta^TQF_zR \quad (16)$$

$$Ev\ \{(\hat{Z}-Z_0)\,(\hat{Z}-Z_0)^T\} = RF_z^TQF_zR - RF_z^TQF_\theta(F_\theta^TQF_\theta)^{-1}F_\theta^TQF_zR \quad (17)$$

and the purely covariance matrices

$$Ev\ \{(\hat{Z}-\tilde{Z})\,(Z_0-\tilde{Z})^T\} = Ev\ \{(\hat{Z}-\tilde{Z})\,(\hat{Z}-\tilde{Z})^T\} \quad (18)$$

$$Ev\ \{(\hat{\theta}-\tilde{\theta})\,(\hat{Z}-\tilde{Z})^T\} = -(F_\theta^TQF_\theta)^{-1}F_\theta^TQF_zR \quad (19)$$

$$Ev\ \{(\hat{\theta}-\tilde{\theta})\,(Z_0-\tilde{Z})^T\} = Ev\ \{(\hat{\theta}-\tilde{\theta})\,(\hat{Z}-\tilde{Z})\} \quad (20)$$

Note from Eqs. (16) and (17), that $Ev\ \{(\hat{Z}-Z_0)\,(\hat{Z}-Z_0)^T\} = R - Ev\ \{(\hat{Z}-\tilde{Z})\,(\hat{Z}-\tilde{Z})^T\}$ and hence that

$$\sigma^2_{\hat{Z}-Z_0} = \sigma^2_{Z_0-\tilde{Z}} - \sigma^2_{\hat{Z}-\tilde{Z}} \quad (21)$$

This result can be used together with the chi-square test described later to spot any gross errors in the measured data by comparing the size of the adjustment, $\hat{Z}-Z_0$, with the estimated standard deviation of the adjustment. The diagonal elements of the matrices calculated in Eqs. (15) and (16) describe the variance of the parameters and of the adjusted data. One finds, in general, that $\sigma^2_{\hat{Z}-\tilde{Z}} < \sigma^2_{Z_0-\tilde{Z}}$. Even though R may be a diagonal matrix, covariance among the adjusted values of Z is introduced as a result of the analysis. This is often an important factor in computing the variance of functions of adjusted data.

For example, in analyzing heat and material balances around a simple distillation column, errors in the distillate rate D and the bottoms rate B will tend to have a negative correlation coefficient. Consequently, errors in the function D/B will often be significantly larger than would be inferred if only the variance of the adjusted values of D and B were taken into account.

USE OF ERROR MATRICES IN EXPERIMENTAL DESIGN

As an example of the use of these matrices in experimental design, consider an adiabatic flash: The feed, F_1, vapor product, F_2, and liquid product, F_3, are each measured in gm/hr units. With these data, only the constraint $F_1-F_2-F_3 = 0$ can be written. If the variance of each measurement is 1.0 gm^2/hr^2 and there is no covariance, we find from Eq. (16),

$$Ev\ \{(\hat{Z}-\tilde{Z})\,(\hat{Z}-\tilde{Z})^T\} = \begin{bmatrix} .67 & .33 & .33 \\ .33 & .67 & -.33 \\ .33 & -.33 & .67 \end{bmatrix}$$

By applying the data analysis techniques, one third of the variance in the flows can be removed.

Now suppose that the temperatures of the three streams can also be measured, each with variance 1.0°C^2, and that there is no covariance between the measurements. If the specific heat of the liquid streams is 0.60 cal/gm°C and the enthalpy of the vapor stream is (39 + .3T_2) cal/gm°C, we can write the heat balance

$$0.6F_1T_1 - (39 + 0.3T_2)\,F_2 - 0.6F_3T_3 = 0$$

In addition, a third constraint, $T_2 - T_3 = 0$, can be written, because streams 2 and 3 are in equilibrium. We find for F_z

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ .6T_1 & -(39+.3T_2) & -.6T_3 & .6F_1 & -.3F_2 & -.6F_3 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

where the columns correspond to F_1, F_2, F_3, T_1, T_2, and T_3. Taking nominal values of these quantities as 10, 8, and 2 gm/hr and 100, 80, and 80°C we find from Eq. (16)

$$Ev\ \{(\tilde{Z}-\hat{Z})\,(\tilde{Z}-\hat{Z})^T\} = \begin{bmatrix} .54 & .44 & .17 & -.11 & .03 & .03 \\ .44 & .45 & -.01 & .21 & -.06 & -.06 \\ .17 & -.01 & .18 & -.32 & .10 & .10 \\ -.11 & .21 & -.32 & .79 & .06 & .06 \\ .03 & -.06 & .10 & .06 & .48 & .48 \\ .03 & -.06 & .10 & .06 & .48 & .48 \end{bmatrix}$$

The columns and rows correspond to F_1, F_2, F_3 T_1 T_2, T_3. The advantage over the previous case is that the variance of the three adjusted flows has been reduced to 0.54, 0.45, and 0.18 vs the values of .67, .67, and .67 when only the mass balance was imposed on the data. Covariance introduced in the adjusted data is evident in both examples. This technique is straightforward, and the additional measurements enhance the accuracy of the adjusted data set.

TEST FOR INTEGRITY OF THE MEASURED DATA

Discrepancies in the constraints as calculated from the measured data reflect error in the measurements themselves. By obtaining the expected distribution of error in the constraints, we can infer whether or not the error in the measured data is consistent with its variance-covariance matrix, R. In this manner, the likely presence of aberrant data can be checked before carrying out the data adjustment calculations. A multivariate chi-square test is proposed for this purpose. Other tests which are helpful in pointing to aberrant data are suggested.

Applying Eq. (4) to the first iteration and rearranging gives

$$HF_zd_1 = -f_0 - HF_\theta(\theta_1-\theta_0) \simeq f(Z_0,\theta_1)$$

As noted previously, $f(Z_0,\theta_1)$ is distributed approximately as $N(0,I)$. Hence the scaler f^Tf has a central chi-square distribution. Eq. (11) is used to evaluate θ, and, as noted in deriving the estimation equation, this vector of parameter values minimizes $f(Z_0,\theta_1)^Tf(Z_0,\theta_1)$. Returning to the original variables, we find $F(Z_0,\theta_1)^TQF(Z_0,\theta_1) \sim \chi^2_{m-p}$. The indicated $m-p$ degrees of freedom of the chi-square statistic arise because p parameters have been estimated from the data.

This statistic can be used to test the null hypothesis that $f(Z_0,\theta_1) \sim N(0,I)$ at a selected level of confidence. If the test is passed, the analysis can proceed with some assurance that the resulting error estimates on the adjusted data, and parameters will either be accurate or conservative.

If the test is failed, the error in the measured variables, Z_0, is inconsistent with the variance-covariance matrix, R, at the level of confidence selected for the test. This can mean either that one or more of the measurements are blunders, or that the variances used for the measured data are too small.

The chi-square test, in addition to being applicable before completing possibly unwarranted data adjustment calculations, is multivariate and is apt to be more powerful than others such as the test on

$$\sum_{j=1}^{n} (\hat{Z}^{(j)} - Z_0{}^{(j)})/\sigma_j \quad \text{(Nogita, 1972)}$$

The m tests on $f(Z_0,\theta_1)$, based on the univariate normal distribution, can also be applied before data adjustment. They are useful in spotting possible blunders if the measured variables affect the discrepancies in a readily discernable pattern. Finally, the n tests based on $(\hat{Z}^{(j)} - Z_0^{(j)})/\sigma_{\hat{Z}_0(j)-Z_0(j)}, j = 1, n$, and the univariate normal distribution, can be used to locate aberrant data. They require that the data adjustment be completed before they can be implemented. The necessary estimates of the variances, $\sigma^2_{\hat{Z}(j)-Z_0^{(j)}}$, can be obtained from Eq. (21).

Should the chi-square test on the measured data fail, it is likely that the adjusted data and parameters will be calculated with error generally greater than indicated by their variance-covariance matrices, although a set of data which is consistent with the constraints can always be obtained. By examining the values of $F^{(j)}/\sigma_{F(j)}$ and possibly $(\hat{Z}^{(j)} - Z_0^{(j)})/\sigma^2_{\hat{Z}(j)-Z_0(j)}$, likely candidates for blunders can be found. These variables can be considered in sequence as parameters, and the chi-square test can be recalculated, with appropriately reduced degrees of freedom. The test can be used as a guide for the suitability of

TABLE 1. VARIABLES IN EXAMPLE PROBLEM

	σ	Measured Value
$Z^{(1)}$	0.017	0.1858
$Z^{(2)}$	0.05	4.7935
$Z^{(3)}$	0.024	1.2295
$Z^{(4)}$	0.2	3.88

TABLE 2. CHI-SQUARE AND UNIT NORMAL TEST RESULTS

Measurement Discarded	χ^2	Unit Normal Test
$Z^{(1)}$	6.8	1.5
$Z^{(2)}$	1.0	0.07
$Z^{(3)}$	1.6	1.2
$Z^{(4)}$	8.4	2.2

ignoring certain measurement values in further calculations with the data set.

We will use the example problem of Ripps (1965) and Nogita (1972) to illustrate the use of the chi-square tests. Table 1 summarizes the necessary data on the measured quantities. There is no covariance in the measured data in this example. The associated constraints are

$$0.1\ Z^{(1)} + 0.6\ Z^{(2)} - 0.2\ Z^{(3)} - 0.7\ Z^{(4)} = 0$$
$$0.8\ Z^{(1)} + 0.1\ Z^{(2)} - 0.2\ Z^{(3)} - 0.1\ Z^{(4)} = 0$$
$$0.1\ Z^{(1)} + 0.3\ Z^{(2)} - 0.6\ Z^{(3)} - 0.2\ Z^{(4)} = 0$$

The chi-square statistic, F^TQF, computed from these data is $\chi^2_3 = 8.3$. The critical value of χ^2_3 at the 95% confidence level is 7.8. Hence a null hypothesis that the standard deviations of Table 1 are consistent with the measured data values is rejected.

Normalizing the balance discrepancies by the standard deviations of the balances as given by $F_zRF_z^T = Q^{-1}$ gives no clear indication in this case of possibly aberrant data. Considering each variable in turn to be a parameter, calculating a "best fit" value for each θ_i using Eq. (5), the values of $F(Z_0, \theta_i)$ and finally $\chi^2_2 = F^T(Z_0, \theta_i)\ QF(Z_0, \theta_i)$ gives the data shown in Table 2. Corresponding values of Nogita's (1972) test statistic, which has a unit normal distribution, are also shown.

At the 95% confidence level, the critical value of χ^2_2 is 5.99, and the values ± 1.96 apply to the unit normal test. The chi-square test shows Z_2 and Z_3 to be the possible gross errors, while the unit normal test includes Z_1, Z_2 and Z_3 as possibly being problematic. There is no statistical justification for examining pairs of variables as possible errors, since two candidates have been uncovered by the analysis.

Measurement 2 has an error of five times its standard deviation in this synthetic problem. From the chi-square test we would, in a real situation, have to turn to an analysis of the experiment itself or examine other data sets to gain more conclusive information about the location of the gross error indicated by the statistics as being in either $Z^{(2)}$ or $Z^{(3)}$.

Independent of which statistical test is applied, $m - p\ (<n)$ facts (constraints) must be used to detect possible gross errors in n measured data values. Experience with other synthetic problems shows that even statistically rare events such as measurement errors of two or three times their standard deviation are not reliably detected by even the chi-square test, although it performs better than univariate tests. The problem of detecting errors becomes more difficult as the ratio of data items to the number of constraints less parameters increases.

In any case, it will be preferable to associate a physical explanation with any indicated aberrant results before discarding data by relegating them to the status of parameters. Statistically rare combinations of measured values could wrongly infer an aberrant value, as could an incorrectly specified variance-covariance matrix of the measured values. Further, the use of an incorrect model as a constraint, such as a second-order kinetic model used on data which, in fact, follow first-order kinetics, could lead to the wrong conclusions about data credibility as well as giving improperly adjusted data values.

NOTATION

B' = $(F_\theta^T Q F_\theta)^{-1} F_\theta^T Q$

B = $B'F_z$

d = reduced adjustment to measured data, n-vector

E' = $RF_z^T Q[I - F_\theta B']$

E = $E'F_z$

Ev = expected value

f = reduced discrepancies in the constraints, n-vector

F = discrepancies in constraint equations, m-vector

F_z = $m \cdot n$ matrix obtained by differentiating the constraint equations with respect to the measured data values

F_θ = $m \cdot p$ matrix obtained by differentiating the constraint equations by the parameters

m = number of constraint equations

n = number of measured data values

p = number of parameters to be estimated from the measured data and constraints

Q = $(F_z R F_z^T)^{-1}$

R = variance-covariance matrix of data values, $n \cdot n$

Z_0 = measured data values, n-vector

Z_i = i th estimate of data values, n-vector

$\hat{Z}$ = final estimate of data values, n-vector

$\tilde{Z}$ = true data values, n-vector

θ_0 = i th estimate of parameter values, p-vector

$\hat{\theta}$ = final estimate of parameter values, p-vector

$\tilde{\theta}$ = true values of parameters, p-vector

Subscript

i = iteration counter, $i=0$ implies measured values or initial estimates in the case of parameters.

Superscript

$H^{(j)}$ = j th element of vector H

LITERATURE CITED

Bennett, C. A. and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*, Wiley and Sons, New York (1954).

Britt, H. I., and R. H. Leucke, "The Estimation of Parameters in Nonlinear Implicit Models," *Technometrics*, **15**, 233 (1973).

Davies, O. L. (Ed), *Statistical Methods in Research and Production*, Hafner Pub. Co., New York (1957).

Demming, W. E., *Statistical Adjustment of Data*, Wiley and Sons, New York (1946).

Draper, N. R. and W. G. Hunter, "Design of Experiments for Parameter Estimation in Multiresponse Situations," *Biometrika* 53, 525-533 (1966).

Himmelblau, D. M., *Process Analysis by Statistical Methods*, Wiley and Sons, New York (1970)

Kuehn, D. R. and H. Davidson, "Computer Control II Mathematics of Control," *Chem. Eng. Prog.*, **57**, 44 (1961).

Lapidus, L., *Digital Computation for Chemical Engineers*, McGraw Hill, New York 249 (1962).

Madron, S., Everka, V. and V. Vanecek, "Statistical Analysis of Material Balance of a Chemical Reactor," *AIChE J.*, **23**, 482 (1977).

Murthy, A. K. S., "A Least Squares Solution to Mass Balance around a Chemical Reactor," *Ind. Eng. Chem. Proc. Des. Dev.*, **12**, 246 (1973).

Nogita, S., "Statistical Test and Adjustment of Process Data," *Ind. Eng. Chem. Proc. Des. Dev.*, **11**, 197 (1972).

Pearson, Carl E., *Handbook of Applied Mathematics*, Van Nostrand Reinhold, New York (1974).

Ripps, D. L., "Adjustment of Experimental Data," *Chem. Eng. Prog. Symposium Series* #55, 8 (1965).

Manuscript received August 14, 1978; revision received May 29, and accepted August 20, 1979.