R & D NOTES

Detection of Gross Errors in Process Data

R. S. H. MAH
and
A. C. TAMHANE

Northwestern University
Evanston, IL 60201

LINEAR RECONCILIATION PROBLEM

Process data reconciliation and rectification and their relationship to process performance monitoring functions have been the subject of many recent publications. [See, for instance, Mah (1981) for a review of these publications.] In this note we shall confine our attention to process data reconciliation subject to linear constraints, and more specifically, to the problem of detecting and identifying the presence of one or more gross errors in the process data.

Generally speaking, process measurements are corrupted by two types of errors: Random errors which are commonly assumed to be independently and normally distributed with zero mean, and gross errors which are caused by non-random events such as instrument biases, malfunctioning measuring devices, incomplete or inaccurate process models. Let $\mathbf{y}$ be an $(n \times 1)$ vector of measured variables, $\mathbf{b}$ be a $(p \times 1)$ vector of unknown parameters, $\mathbf{D}$ be an $(n \times p)$ matrix of known constants, for which rank $(\mathbf{D}) = p \leq n$, and $\boldsymbol{\epsilon}$ be an $(n \times 1)$ vector of errors distributed normally with a zero mean vector and a known variance-covariance matrix Q. Then in the absence of gross errors, the basic model is

$$\mathbf{y} = \mathbf{Db} + \boldsymbol{\epsilon} \tag{1}$$

and the general linear reconciliation problem is the least-squares estimation of $\mathbf{b}$ subject to the linear constraints

$$\mathbf{Ab} = \mathbf{c} \tag{2}$$

where $\mathbf{A}$ is a $(q \times p)$ matrix of known constants and $\mathbf{c}$ is a $(q \times 1)$ vector of known constants.

The linear reconciliation problem formulated above is a generalization of the reconciliation problems treated by previous investigators. Thus, the reconciliation of flow and inventory data reported by Mah et al. (1976) is a special case in which $\mathbf{y}$ is the vector of measured flow rates ($\mathbf{v}$ in their paper), $\mathbf{D}$ is an identity matrix, $\mathbf{b}$ is the vector of true flow rates $\mathbf{s}(\boldsymbol{\mu})$, $\mathbf{A}$ is the incidence matrix ($\mathbf{A}$), $p = n$ and $\mathbf{c} = \mathbf{0}$. Nogita (1972) treated essentially the same problem but considered only the diagonal terms (variances) of the covariance matrix in his minimization. Almasy and Sztano (1975) also studied this problem but they allowed $\mathbf{c}$ to be non-zero. On the other hand, the reactor data reconciliation problem reported by Madron et al. (1977) contains no constraints (Eq. 2) on $\mathbf{b}$ which corresponds to the vector of extents of chemical reactions ($\boldsymbol{\xi}$). For that problem $\mathbf{y}$ is the measured vector of increases in the numbers of moles of species ($\mathbf{n}^+$), $\mathbf{D}$ is their $(J \times I)$ matrix ($\mathbf{A}^T$) of stoichiometric coefficients, $n = I$ = number of reactive species, J = number of independent chemical reactions, and $\mathbf{Q}$ is denoted by $\underline{F}$ in their paper. Madron et al. (1977) actually considered an r-subvector of $\mathbf{n}^+$ (denoted by $\mathbf{n}_r^+$ in their paper) corresponding to the $r \leq I$ species for which measurements were made. A similar problem was studied by Murthy (1973, 1974).

Notice in the general linear reconciliation problem formulated above we have tacitly assumed all variables to be measured. If this assumption is not true, we can always perform node aggregation (Mah et al., 1976) or appropriate output assignment and sorting (Romagnoli and Stephanopoulos, 1980) to obtain a reconciliation problem of lower dimension. We can therefore make this assumption without any loss of generality.

For the above formulation the least-squares estimate of $\mathbf{b}$, which is also the maximum likelihood estimate and also the minimum variance unbiased estimate, is (Seber, 1977, p. 85, Eqs. 3–59)

$$\hat{\mathbf{b}} = \hat{\mathbf{b}}_0 + (\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{D})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{D})^{-1}\mathbf{A}^T]^{-1}(\mathbf{c} - \mathbf{A}\hat{\mathbf{b}}_0) \tag{3}$$

where $\hat{\mathbf{b}}_0$ is the unconstrained least-squares estimate of $\mathbf{b}$ given by

$$\hat{\mathbf{b}}_0 = (\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{y}. \tag{4}$$

For the special case considered by Mah et al. (1976) we get

$$\hat{\mathbf{b}} = [\mathbf{I} - \mathbf{QA}^T(\mathbf{AQA}^T)^{-1}\mathbf{A}]\mathbf{y} \tag{5}$$

which is their Eq. (5). Similarly, with appropriate change of notation Eq. 4 becomes Eq. 14 of Madron et al. (1977).

Let $\hat{\mathbf{y}} = \mathbf{D}\hat{\mathbf{b}}$ where $\hat{\mathbf{b}}$ is given by Eq. (3). We shall refer to $\hat{\mathbf{y}}$ as the vector of adjusted or smoothed measurements.

DETECTION OF GROSS ERRORS

Data reconciliation deals with the problem of random errors. If gross errors are also present in the process data, they must be identified and removed (by discarding the corresponding measurements) before reconciliation. In this paper we shall consider the gross errors to be associated with the measurements rather than the process model. The question to be answered is whether a gross error is present in at least one of the y_i's, and if yes, which ones?

A simple test to answer this question can be based on the residuals,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{DM})\mathbf{y} - \mathbf{DNc} \tag{6}$$

where

$$\mathbf{M} = (\mathbf{I} - \mathbf{NA})(\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Q}^{-1} \tag{7}$$

and

$$\mathbf{N} = (\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{D})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{D}^T\mathbf{Q}^{-1}\mathbf{D})^{-1}\mathbf{A}^T]^{-1}. \tag{8}$$

It is easy to show that

$$\mathbf{e} \sim N(\mathbf{0},\mathbf{V}) \tag{9}$$

where

$$\mathbf{V} = (\mathbf{I} - \mathbf{DM})\mathbf{Q}(\mathbf{I} - \mathbf{DM})^T. \tag{10}$$

Therefore, under H_{oi}: There is no gross error in the ith observation,

Correspondence concerning this paper should be addressed to Prof. R. S. H. Mah.
0001-1541-82-6427-0828-$2.00.

$$e_i \sim N(0, v_{ii}). \tag{11}$$

We would reject H_{oi} and conclude that there is a gross error present in the ith observation, if

$$|Z_i'| = |e_i|/\sqrt{v_{ii}} > Z_{\alpha/2} \tag{12}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ point [i.e., $(1-\alpha/2)$th quantile] of the standard normal distribution and α is the level of significance.

Very recently, Tamhane (1981) has shown that a better test may be based on the transformed residual vector

$$\mathbf{d} = \mathbf{Q}^{-1}\mathbf{e} \sim N(\mathbf{0}, \mathbf{W}) \tag{13}$$

where

$$\mathbf{W} = \mathbf{Q}^{-1}\mathbf{V}(\mathbf{Q}^{-1})^{\mathrm{T}}. \tag{14}$$

It can be shown that the test statistics

$$Z_i = d_i/\sqrt{w_{ii}} \tag{15}$$

possesses the maximal power for detecting the presence of a single outlier. [For a review of statistical literature on detection of outliers, see Barnett and Lewis (1978).] We would conclude that the ith observation is an outlier (contains a gross error) with a type I error probability of α, if

$$|Z_i| > Z_{\alpha/2}. \tag{16}$$

The above test may be modified to take into account of the effect of carrying out multiple tests ($i = 1, 2, \ldots, n$). The idea is that for multiple tests, all at level α, the probability of rejecting at least one H_{oi} when in fact all H_{oi} are true (i.e., the probability of overall type I error) can be much larger than α. It approaches one, as n tends to infinity. To account for this effect and to guarantee that the probability of overall type I error is controlled at α, the test Eq. (16) may be modified in the following way: Reject H_{oi} and conclude that there is a gross error present in the ith observation, if

$$|Z_i| > Z_{\beta/2} \tag{17}$$

where

$$\beta = 1 - (1-\alpha)^{1/n}. \tag{18}$$

Since $\beta < \alpha$ for $n > 1$, $Z_{\beta/2} > Z_{\beta/2}$ making it more difficult to reject H_{oi}.

In fact, even in the situation where a single gross error is suspected but the source of the gross error is unknown, it is appropriate to use the critical point $Z_{\beta/2}$ and not $Z_{\alpha/2}$. This is so because in this situation the test should be based on $|Z|_{\max} = \max_{1\le i\le n} |Z_i|$ and a gross error in the observation corresponding to $|Z|_{\max}$ is indicated, if $|Z|_{\max}$ exceeds the critical value. The null distribution of $|Z|_{\max}$ is not standard normal and hence $Z_{\alpha/2}$ is not the appropriate critical value for guaranteeing type I error probability of α. The null distribution of $|Z|_{\max}$ is in general complicated but it can be shown (Sidak, 1967) that $Z_{\beta/2}$ provides an upper bound on the exact α upper point of $|Z|_{\max}$ and thus a conservative test.

POWER OF THE PROPOSED TEST

When gross errors are present, $E(\underline{\epsilon}) = \underline{\delta}$, where $E(\epsilon_i) = \delta_i \neq 0$, if gross error is present in the ith observation. Now

$$\begin{aligned} E(\mathbf{d}) &= E(Q^{-1}\mathbf{e}) = Q^{-1}E[(\mathbf{I}-\mathbf{DM})\mathbf{y} - \mathbf{DNc}] \\ &= Q^{-1}[(\mathbf{I}-\mathbf{DM})(\mathbf{Db} + \underline{\delta}) - \mathbf{DNc}] \\ &= Q^{-1}(\mathbf{I}-\mathbf{DM})\boldsymbol{\delta} = \mathbf{G}\,\boldsymbol{\delta}. \end{aligned} \tag{19}$$

If only one gross error is present, say, in the nth observation, then

$$E(Z_i) = E(d_i)/\sqrt{w_{ii}} = \delta_n g_{in}/\sqrt{w_{ii}} = \nu_i,\ i = 1, 2, \ldots, n \tag{20}$$

and the power of the test is bounded above by

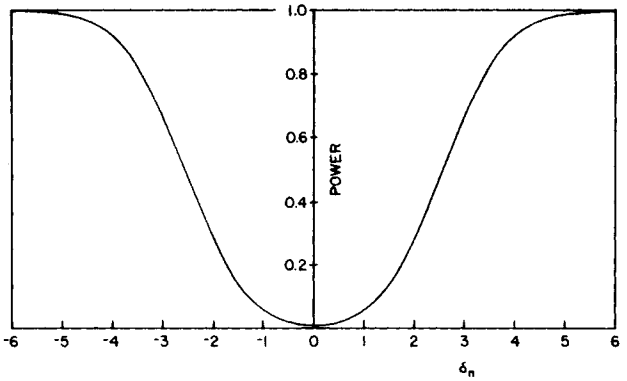

Figure 1. Power of the test for $\nu_n = \delta_n$, $\alpha = 0.05$.

$$\begin{aligned} P\{|Z_n| > k\} &= 1 - P\{-k \le Z_n \le k\} \\ &= 1 - P\{-k-\nu_n \le Z_n - \nu_n \le k - \nu_n\} \\ &= 1 - \{\Phi(k-\nu_n) - \Phi(-k-\nu_n)\} \\ &= \Phi(\nu_n - k) + \Phi(-\nu_n - k) \end{aligned}$$

where

$$k = Z_{\beta/2} \tag{22}$$

and

$$\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2}\exp(-t^2/2)\,dt \tag{23}$$

is the standard normal distribution function. Figure 1 shows that as the magnitude of the gross error increases, so does the probability of its detection. Strictly speaking, the power should be written as $P\{|Z_n| > |Z_i|\ (1 \le i \le n-1), |Z_n| > k\}$. But as δ_n increases, it becomes practically the same as Eq. (21). For the general situation involving possible presence of multiple gross errors, computer simulation may be used to obtain an indication of the power of the test.

DISCUSSION

We note that rank (V) = rank (W) = $n - (p - q)$, where q = rank (A) $\le p$. The estimates $\hat{\mathbf{y}}$ lie in the $(p-q)$ dimensional subspace of R^n, whereas the residuals e lie in its $(n-p+q)$ dimensional orthogonal complement. There is a trade-off between the estimation space which contains the reconciled estimates of parameters (state variables) and the residual space which contains the information for gross error detection. If the constraints, Eq. (2), are totally absent, and rank (D) = n, e will lie in a nullspace, and the measurements will be perfectly fitted. In other words, all the information derived from measurements will be used for estimation and none for gross error detection. In the other extreme, if rank (A) = n, then b and $\hat{\mathbf{y}}$ may be estimated without using measurements, and rank (V) = rank (W) = n. In that case all the information derived from measurements will be available for gross error detection, and none will be used for estimation.

In general, the rank of V or W will be less than n, and it is possible to have the same Z_i for different measurements, making it impossible to differentiate between them in gross error identification. This problem will arise if we allow more than two streams linking the same pair of nodes in the process digraph, which introduces an inherent nonidentifiability unless it is augmented with further information, for instance, stoichiometric information. If the number of distinct values of Z_i, say m, is less than n, a less conservative multiple test may be obtained by using m in place of n in Eq. (18).

In this note we have shown the reconciliation of process flow and reactor data as special cases of least-squares estimation with and

without linear constraints and the detection of gross errors in process data as a problem of testing for outliers in statistical data. The proposed gross error detection scheme makes use of a recently-developed test which possesses maximal power properties for detecting the presence of a single outlier (gross error). This test can be extended to deal with multiple outliers but without the guarantee of maximal power properties. The procedure which is applicable to any linear reconciliation problem is simple to apply and to program on computers.

For process data obeying normal distribution but containing a single gross error the power of this test is given in closed analytical form. Its properties for a more realistic situation may be obtained by computer simulation.

By dealing directly with the residuals (the differences between observed and fitted values) the necessity of interposing an identification scheme following the detection of one or more gross errors is eliminated. Once a gross error is detected, its origin is automatically identified.

Acknowledgment

Partial support of this work was provided by the National Science Foundation Grant CPE 76-18852.

NOTATION

$\mathbf{A}$ = a $(q \times p)$ matrix of known constants
$\mathbf{b}$ = a $(p \times 1)$ vector of unknown parameters
$\hat{\mathbf{b}}$ = least-squares estimate of $\underline{b}$
$\hat{\mathbf{b}}_0$ = the unconstrained least-squares estimate of $\underline{b}$
$\mathbf{c}$ = a $(q \times 1)$ vector of known constants
$\mathbf{d}$ = a transformed residual vector defined by Eq. (13)
$\mathbf{D}$ = an $(n \times p)$ matrix of known constants
$\mathbf{e}$ = the residuals $\underline{y} - \hat{\underline{y}}$
$E(\cdot)$ = expected value of
$\mathbf{G}$ = $\underline{Q}^{-1}(\underline{I} - \underline{DM})$
$\mathbf{I}$ = identity matrix
k = $Z_{\beta/2}$, the $(1 - \beta/2)$th quantile of the standard normal distribution
$\mathbf{M}$ = a $(p \times n)$ matrix defined by Eq. (7)
n = the number of measured variables
$\mathbf{N}$ = a $(p \times q)$ matrix defined by Eq. (8)
p = the number of unknown parameters
$P\{\cdot\}$ = probability of
q = the number of linear constraints
$\mathbf{Q}$ = an $(n \times n)$ variance-covariance matrix
$\mathbf{V}$ = variance-covariance matrix of the residuals, $\underline{e}$
$\mathbf{W}$ = variance-covariance matrix of transformed residuals, $\underline{d}$
$\mathbf{y}$ = an $(n \times 1)$ vector of measured variables
$\hat{\mathbf{y}}$ = an $(n \times 1)$ vector of adjusted measurements
Z_i = a test statistic defined by Eq. (15)
$|Z_i|$ = a test statistic defined by Eq. (12)
α = level of significance
β = modified level of significance defined by Eq. (18)
δ_i = a gross error
ϵ = an $(n \times 1)$ vector of errors
ν_i = the expected value of Z
Φ = standard normal distribution function, Eq. (23)

LITERATURE CITED

Almasy, G. A., and T. Sztano, "Problems of Control and Information Theory," **4**, (1), 57 (1975).
Barnett, V., and T. Lewis, *Outliers in Statistical Data*, John Wiley, New York (1978).
Madron, F., V. Veverka and V. Vanecek, *AICHE J.*, **23**, 482 (1977).
Mah, R. S. H., "Design and Analysis of Process Performance Monitoring Systems," Proceedings of the Engineering Foundation Conference on "Chemical Process Control II," Sea Island, GA, (Jan. 18–23, 1981).
Mah, R. S. H., G. M. Stanley, and D. M. Downing, *Ind. Eng. Chem. Proc. Des. Dev.*, **15**, 175 (1976).
Murthy, A. K. S., *Ind. Eng. Chem. Proc. Des. Dev.*, **12**, 246 (1973).
Murthy, A. K. S., *ibid.*, **13**, 347 (1974).
Nogita, S., *Ind. Eng. Chem. Proc. Des. Dev.*, **11**, 197 (1972).
Romagnoli, J. A., and G. Stephanopoulos, *Chem. Eng. Sci.*, **35**, 1067 (1980).
Seber, G. A. F., *Linear Regression Analysis*, John Wiley, New York (1977).
Sidak, Z., *J. Amer. Statist. Assoc.*, **62**, 626 (1967)
Tamhane, A. C., "A Note on the Use of Residuals for Detecting an Outlier in Linear Regression," to appear in *Biometrika*.

Manuscript received August 10, 1981; revision received December 11, and accepted January 13, 1982.

Laminar Flow in the Entrance Region of a Parallel Plate Channel

A. K. MOHANTY and REETA DAS

Department of Mechanical Engineering
Indian Institute of Technology
Kharagpur, India

Fluid flow in the entrance region of a channel, or a pipe is characterized by non-similar velocity profiles. Typically, a length of 150 diameters may exist in a laminar flow at Reynolds number equal to 2,000, before the fully developed Poiseuille profile is established. It is quite possible, therefore, that in a large majority of practical applications, such as in connecting pieces or in heat exchangers, of much interest to chemical engineers, the transport phenomena are confined to the entrance region.

The problem of laminar flow in the entrance region of pipes and ducts has been studied extensively by several investigators, e.g., Schiller (1922), Schlichting (1979), Wang and Longwell (1964). These investigations were based, primarily, on the assumption that fully developed velocity profile is established at the location where the boundary layers meet at the duct axis. However, in a recent study of flow through a circular pipe (Mohanty and Asthana, 1979), it has been shown, both analytically and experimentally, that the boundary layers meet much earlier and the velocity profile undergoes adjustment in a purely viscous region to finally attain the fully developed form. The boundary layer region is called the "inlet region," and the viscous zone the "filled region," after Shingo (1966).

The present study is aimed at extending the inlet and filled region model to a parallel plate channel.