Data Reconciliation and Gross-Error Detection for Dynamic Systems

João S. Albuquerque and Lorenz T. Biegler
Dept. of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15232

Gross-error detection plays a vital role in parameter estimation and data reconciliation for dynamic and steady-state systems. Data errors due to miscalibrated or faulty sensors or just random events nonrepresentative of the underlying statistical distribution can induce heavy biases in parameter estimates and reconciled data. Robust estimators and exploratory statistical methods for the detection of gross errors as the data reconciliation is performed are discussed. These methods have the property insensitive to departures from ideal statistical distributions and to the presence of outliers. Once the regression is done, the outliers can be detected readily by using exploratory statistical techniques. Optimization algorithm and reconciled data offer the ability to classify variables according to their observability and redundancy properties. Here an observable variable is an unmeasured quantity that can be estimated from the measured variables through the physical model, while a nonredundant variable is a measured variable that cannot be estimated other than through its measurement. Variable classification can be used to help design instrumentation schemes. An efficient method for this classification of dynamic systems is developed. Variable classification and gross-error detection have important connections, and gross-error detection on nonredundant variables has to be performed with caution.

Introduction

While data reconciliation is a common tool for steady-state systems, application to dynamic systems is still in its infancy (Liebman et al., 1992; Sistu et al., 1993). Moreover, the dynamic problem has some interesting characteristics related to the optimization formulation and interpretation of results. In a previous article (Albuquerque and Biegler, 1996) we developed an efficient successive-quadratic-programming-based (SQP) method for data reconciliation and parameter estimation for systems described by differential and algebraic equation systems (DAE). Here the problem is posed as minimizing an objective function that reflects the distributional structure of the measurement errors subject to the process model. The most commonly used objective function is least squares, which assumes that the measurement errors are independent with time and normally distributed. If these assumptions are verified, then the least-squares function leads to the optimal unbiased estimator, in the sense that of all the possible unbiased estimators, it is the one with the smallest variance.

Correspondence concerning this article should be addressed to L. T. Biegler.

However, if the measurement errors do not follow these ideal assumptions (e.g., they are not normally distributed) then the estimates will be biased. That is the case when outliers are present in the data. These can be caused either by a malfunction on a measuring device (e.g., a broken wire in a thermocouple), by improper use of such devices (e.g., a flowmeter ring being installed backwards), or any number of random causes. By definition, an outlier or a gross error is a measurement in which the error does not follow the statistical distribution of the bulk of the data. A number of approaches have been proposed to deal with this event in steady-state systems. Initially, sequential tests were used where a regression using the least-squares assumption was performed, and the residuals were studied using statistical tests in order to see which of those residuals followed the initial distributional assumptions (Narasimhan and Mah, 1988; Tamhane et al., 1992). Once a gross error was detected, a new regression was performed and tests were done until the data were free of errors. This approach has the disadvantage in that the distributional tests are based on the residuals

from a regression that may be heavily biased by the presence of the outliers. An outlier in one measurement may show up in a totally different measurement after the regression is performed. These are defined as leverage points (Rawlings, 1988). Also due to the sequential nature of the method, many regressions may have to be performed until we are satisfied.

Another approach is to take into account the presence of outliers from the very beginning. A common method is the use of contaminated error distributions (Jeffreys, 1932; Fariss and Law, 1979; Tjoa and Biegler, 1991), in which the objective function allows for two error structures, each with a certain probability of occurrence. In this manner, the regression accommodates the presence of outliers, and gross-error detection can be performed simultaneously. However, this approach will not work well if the gross-error distribution is not well characterized. Also it often leads to nonconvex and complex objective functions prone to underflow problems. Another simultaneous approach is to use objective functions that, due to their mathematical structure, are insensitive to deviations from the ideal assumptions (Rey, 1983; Huber, 1980). These estimators tend to look at the bulk of the data and ignore atypical values. In this fashion, an accurate regression can be performed even if nothing is known about the outliers or even the error structure of the data, and exploratory methods can be used to detect gross errors and derive more information from the statistical properties of the data once the residuals are estimated (Tukey et al., 1983).

Finally, uniqueness of the reconciled measurements is strongly linked to the problem formulation, the performance of the optimization, and statistical interpretation of the results. This must be analyzed by a careful variable classification. For steady-state processes (Crowe, 1989; Stanley and Mah, 1981; Kretsovalis and Mah, 1988; Swartz, 1989) developed methods based on necessary and sufficient conditions for redundancy and observability. We extend this work using efficient sparse linear algebra methods and introduce the concept of *collective redundancy*. This concept applies especially to dynamic systems as well as to the more general situation when several data sets are considered together.

In the next section we describe and compare the Bayesian and the robust approaches for gross-error detection. In the third section we address the question of variable classification, and in the fourth section we further consider robust and exploratory statistical methods and how they relate to variable classification. Finally, we illustrate the use and effectiveness of these techniques on a few examples in the fifth section, and conclude this article with some remarks in the last section.

Simultaneous Gross-Error Detection

Consider the following error-in-all-variable-measurements (EVM) constrained by a DAE and separable boundary conditions:

$$\begin{aligned} &\min \sum_{i=1}^{N} l_i(z_i, \bar{z}_i, u_i, \bar{u}_i, \theta) + l_{N+1}(z_{N+1}, \bar{z}_{N+1}, \theta) \\ &\text{s.t.} \quad \Psi\left(\frac{dz}{dt}, z, u, \theta, t\right) = 0 \\ &R_I z(t_1) = z_I, \quad R_T z(t_{N+1}) = z_T \\ &h(\theta) \le 0. \end{aligned} \tag{1}$$

Here we assume that the measured outputs are subsets of the state variable vector, without loss of generality; z_i and u_i are the state and input variables at time $t = t_i$, respectively; $\bar{z}_i$ and $\bar{u}_i$ are their respective measurements; and θ are the unknown parameters. As in our previous approach, we discretize the DAE model using standard implicit Runge-Kutta methods (IRK) (Brenan et al., 1989). The structure of these equations is explored further in the next section.

In this section we describe and compare two approaches where the objective function of problem 1 is constructed in such a way that gross-error detection can be performed simultaneously: the Bayesian approach where the posterior distribution function takes into account the presence of outliers by describing their error structure, and the robust approach where the estimators are designed so that they are insensitive to outliers. We discuss the advantages and disadvantages of each approach.

Bayesian approach

Here we group the state variables z_i and the input variables u_i into $x_i^T = (z_i^T, u_i^T)^T$ and further partition x_i into the measured variables x_i^m and unmeasured variables x_i^u, such that $x_i^T = [(x_i^m)^T, (x_i^u)^T]^T$. We also group the measured and unmeasured variables across the time instants $(x^m)^T = [(x_1^m)^T, \ldots, (x_{N+1}^m)^T]^T$, $(x^u)^T = [(x_1^u)^T, \ldots, (x_{N+1}^u)^T]^T$, and we consider the following error model for the measured variables:

$$\bar{x}^m = x^m + \epsilon, \tag{2}$$

where ϵ is the measurement error. Also, suppose that the variables x^m and x^u are constrained by some physical model and are dependent on unknown parameters θ. Here we represent the equalities in Eq. 1 by the general relation:

$$f(x^m, x^u, \theta) = 0. \tag{3}$$

The problem here is of estimating θ and x^u knowing the measurement $\bar{x}^m$ and the probability distribution function of ϵ. Let $p(\epsilon)$ be the error distribution. Then the measurements $\bar{x}^m$ will be distributed according to $p(\bar{x}^m - x^m | \theta, x^u)$. This will be the likelihood function $L(\theta)$. According to Bayes' theorem, the posterior distribution of the parameters and the unmeasured variables, given the data, will be

$$p(\theta, x^u | \bar{x}^m) = L(\theta, x^u)\Pi(\theta, x^u), \tag{4}$$

where $\Pi(\theta, x^u)$ is the *a priori* distribution function of the parameters θ and the unmeasured variables x^u. Using the maximum *a posteriori* method, we minimize the negative of the log posterior subject to the model:

$$\begin{aligned} &\min_{\theta, x^u} -\log p(\theta, x^u | \bar{x}^m) \\ &\text{s.t.} \quad f(x^m, x^u, \theta) = 0. \end{aligned} \tag{5}$$

If the measurement noise ϵ is normally distributed and independent across the data sets, and if we use a flat prior $\Pi(\theta, x^u)$, problem 5 will become a nonlinear least-squares problem.

To take into account the presence of outliers, we redevelop the contaminated normal distribution (Jeffreys, 1932; Fariss and Law, 1979; Tjoa and Biegler, 1991) as the outcome of a Bernoulli trial (DeGroot, 1986). Here the two possible outcomes are $G = \{Gross\ error\ occurred\}$ with probability η and $R = \{random\ error\ occurred\}$ with probability $1-\eta$. When G occurs, the error ϵ will follow a distribution $p(\epsilon|G)$. This will be the gross-error distribution. Likewise, when event R is true, then the error ϵ will follow $p(\epsilon|R)$. Thus the distribution of ϵ will be

$$p(\epsilon) = (1-\eta)p(\epsilon|R) + \eta p(\epsilon|G). \tag{6}$$

The posterior distribution on θ and x^u will then be

$$p(\theta, x^u|\bar{x}^m) = [(1-\eta)p(\bar{x}^m - x^m|\theta, x^u, R) + \eta p(\bar{x}^m - x^m|\theta, x^u, G)]\Pi(\theta) \tag{7}$$

and this distribution will be used on problem 5 leading to

$$\min_\theta -\log\{[(1-\eta)p(\bar{x}^m - x^m|\theta, x^u, R) + \eta p(\bar{x}^m - x^m|\theta, x^u, G)]\Pi(\theta, x^u)\}$$
$$\text{s.t. } f(x^m, x^u, \theta) = 0. \tag{8}$$

Once problem 8 is solved, gross errors can be identified when the gross-error term is greater than the random-error term for a particular measurement:

$$\eta p(\bar{x}^m - x^m|\theta, x^u, G) > (1-\eta)p(\bar{x}^m - x^m|\theta, x^u, R). \tag{9}$$

Commonly used distributions for the random and gross errors are normal distributions in which the gross-error distribution has a higher variance than the random-error distribution, and where the individual measurements are independent. With a flat uniform prior this leads to the η-contaminated normal distribution with normal contamination (Jeffreys, 1932; Fariss and Law, 1979; Tjoa and Biegler, 1991; Johnston and Kramer, 1995):

$$p(\theta, x^u|\bar{x}^m) \quad \alpha\,(1-\eta)\exp\left\{-\frac{1}{2}\left(\frac{x^m - \bar{x}^m}{\sigma}\right)^2\right\} + \frac{\eta}{b}\exp\left\{-\frac{1}{2}\left(\frac{x^m - \bar{x}^m}{b\sigma}\right)^2\right\}, \tag{10}$$

where $b > 1$ is the ratio of the standard deviation of the gross-error distribution to the standard deviation of the random-error distribution. The contaminated normal distribution has heavier tails than the pure normal distribution, making it less sensitive to outliers. For the same reason, some authors prefer to use a t distribution to describe the gross errors (Verdinelli and Wasserman, 1991). The advantages of the Bayesian approach are that it incorporates knowledge of the parameters and the error structure and allows for straightforward inferencing because of its Bayesian nature. However, if not much is known about the gross-error structure, or even the random-error structure, then this approach can be very misleading. Also, the posterior distributions, Eq. 7, have a complicated mathematical structure, prone to numerical problems such as underflows, and they lead to nonconvex objective functions in problem 8, making this problem more difficult to solve. This is examined in more detail later in this section.

Robust estimators

In the previous section we discussed some of the problems that plague the Bayesian-based gross-error detection methods. The use of robust estimators alleviates these problems, though introducing some of its own. With the classic approach we assume that the measurement errors follow a certain statistical distribution, and all statistical inferences are based on that distribution. However, departures from all ideal distributions (such as outliers) can invalidate these inferences. In robust statistics, rather than assuming an ideal distribution, we construct an estimator that will give unbiased results in the presence of this ideal distribution, but that will be insensitive to deviations from ideality to a certain degree. Suppose that $\{\xi_1, \ldots, \xi_n\}$ are drawn from distribution $f(\xi)$ and let T be an unbiased estimator $\hat{\theta} = T[f(\xi)]$ of parameter θ. If we have only a more or less valid distributional model $g(\xi)$, then the estimate will be $\tilde{\theta} = T[g(\xi)]$. The distributions of the estimators based on $f(\xi)$ and $g(\xi)$ will be $\Gamma(\hat{\theta}, f)$ and $\Gamma(\tilde{\theta}, g)$. The estimator $T(\cdot)$ will be *robust* iff

$$d(f,g) < \eta \Rightarrow d[\Gamma(\hat{\theta}, f), \Gamma(\tilde{\theta}, g)] < \epsilon, \tag{11}$$

where $d(\cdot)$ is a distance function. Thus a bounded shift from the ideal assumptions will lead to a bounded shift in the estimates. To assess robustness, an important notion is the influence function. The influence function measures the importance of an observation on the estimator and is defined as:

$$\Psi(\xi_0) = \lim_{t \to 0} \frac{T[(1-t)f + t\delta(\xi - \xi_0)] - T[f]}{t}, \tag{12}$$

where $\delta(\xi - \xi_0)$ is the Dirac function centered on a particular observation ξ_0. For an estimator to be robust, its influence function has to be bounded as the observation ξ_0 is taken to infinity.

M-estimators: fair function

There are several classes of robust estimators. The most important for us are the M-estimators, which are generalizations of the maximum-likelihood estimator. Their form is

$$\min_\theta \sum_{i=1}^{n} \rho(\xi_i, \theta). \tag{13}$$

Several estimates have been proposed in the literature (Rey, 1983; Huber, 1980; Basu and Paliwal, 1989), but to our knowledge only the fair function is convex and has continuous first and second derivatives. Its form is

$$\rho(\xi) = c^2\left[\frac{|\xi|}{c} - \log\left(1 + \frac{|\xi|}{c}\right)\right], \tag{14}$$

where c is a tuning parameter. Another important notion is the *relative efficiency* for some nominal distribution, say Gaussian. For the single-parameter case, it is defined as

$$E = \frac{V_{opt}}{V_{act}}, \tag{15}$$

where V_{opt} is the optimal error variance of an estimator with this nominal distribution given by the Cramer–Rao inequality (DeGroot, 1986) and V_{act} is the error variance attained by the estimator considered. As c becomes smaller, the fair function estimator becomes less sensitive to outliers and its influence function becomes smaller. However its asymptotic efficiency with respect to the normal distribution also decreases. If E is the relative asymptotic efficiency with the normal distribution, for the single parameter case, then c is crudely related to it by (Rey, 1983)

$$c = 0.21529\left(\frac{E - 0.63662}{1 - E}\right)^{1.02}. \tag{16}$$

In general, there is a trade-off between robustness and efficiency. The more robust an estimator is, the less efficient it is. For M estimators, the influence function is proportional to the estimator's derivative. For a least-squares estimator, the influence function would be

$$\Psi_{ls}(\xi) \propto \xi. \tag{17}$$

For the contaminated normal, we take the derivative of the $-\log$ of Eq. 10:

$$\psi_{cont}(\xi) \propto w(\xi)\xi, \tag{18}$$

where $w(\xi)$ is a weighting function given by

$$w(\xi) = \frac{(1-\eta)\exp\left\{-\frac{1}{2}\left(\frac{\xi}{\sigma}\right)^2\right\} + \frac{\eta}{b}\exp\left\{-\frac{1}{2}\left(\frac{\xi}{b\sigma}\right)^2\right\}}{(1-\eta)\exp\left\{-\frac{1}{2}\left(\frac{\xi}{\sigma}\right)^2\right\} + \frac{\eta}{b^3}\exp\left\{-\frac{1}{2}\left(\frac{\xi}{b\sigma}\right)^2\right\}}. \tag{19}$$

For very large values of ξ, this weighting function can be approximated by $w(\xi) \approx 1/b^2$, whereas for small values of ξ, $w(\xi) \approx 1$. Therefore, a good approximation to the influence function of the contaminated normal estimator will be

$$\Psi_{cont}(\xi) \propto \begin{cases} \xi, & \xi \to 0 \\ \frac{1}{b^2}\xi, & \xi \to \infty. \end{cases} \tag{20}$$

The influence function for the fair function is given by

$$\Psi_{fair}(\xi) \propto \frac{\xi}{1 + \frac{|\xi|}{c}}. \tag{21}$$

Taking a very large observation, $\xi \to \infty$, the influence function for both the least-squares and the contaminated normal will grow to ∞, whereas the influence function for the fair estimator will be bounded since $\lim_{\xi \to \infty} \Psi_{fair}(\xi) \propto c$. Note that the Bayesian-type estimators described earlier normally do not have a bounded influence function, because for large residuals the posterior distribution 7 is approximated by the gross-error distribution (say a normal or a t distribution). A maximum *a posteriori* estimator based on this distribution will not be robust, except in the special case when the Cauchy distribution (DeGroot, 1986) is used to model the gross errors. However this distribution does not lead to a convex function.

The fair estimator is not scale invariant, so if we are trying to estimate x from its measurement $\bar{x}$, we should use an estimate of the scale σ. The fair function would then be

$$\rho(\epsilon, c) = c^2\left[\frac{|\epsilon|}{c} + \log\left(1 + \frac{|\epsilon|}{c}\right)\right], \tag{22}$$

with $\epsilon = (x - \bar{x})/\sigma$ being the estimated residual. It is sometimes also convenient to multiply Eq. 22 by σ^2 to scale it up. This avoids numerical problems such as very high values of the objective function and gradients, and small values for the second derivatives. This estimator behaves like the least-squares estimator for small residuals, but like an absolute-value estimator for large residuals. For small residuals we take a second-order Taylor expansion around zero. For large magnitude residuals, the logarithmic term will be much smaller than the absolute-value term. For these extremes, the fair function, Eq. 22, behaves like

$$\rho(\epsilon, c) = \begin{cases} \frac{1}{2}\epsilon^2 & \text{for small } \epsilon \\ c|\epsilon| & \text{for large } \epsilon. \end{cases} \tag{23}$$

This robust estimator has the advantage of having a very simple mathematical form and of having very convenient properties for optimization. It also does not require any knowledge of the error structure of the outliers and of the data themselves. However, it cannot be used to draw inferences to the same extent as the Bayesian estimators because we are not incorporating any knowledge of the distributional structure of the data, but rather using a mathematical construct that is resistant to outliers.

Objective function properties for data reconciliation

The properties of the first and second derivatives of the estimator are important for the robustness and efficiency of the data-reconciliation procedure. Consider the following simplified problem:

$$\begin{aligned} &\min \phi(x) \\ &\text{s.t. } f(x) = 0. \end{aligned} \tag{24}$$

The Lagrange function for problem 24 is given by

$$L(x) = \phi(x) + \gamma^T f(x), \tag{25}$$

where γ are the multipliers. First-order optimality conditions require that the multipliers lie in the range space of the linearized constraints at the solution x^*:

$$\gamma = -(Y^T A)^{-1} Y^T g, \tag{26}$$

where A is the Jacobian of the constraints $f(x)$ evaluated at x^*; Y is some matrix whose columns span the range space of A; and g is the gradient of the objective function $\phi(x)$ at x^*. Now suppose that the residuals in either the log posterior of the contaminated normal function 10 or the fair estimator 22 are small. The gradient for the log posterior of the contaminated normal is given by Eqs. 18 and 19, and it is easy to see that it will be small for small residuals. For the fair function, Eq. 22, the gradient will be given by

$$\frac{\partial \rho(\epsilon, c)}{\partial \epsilon} = \frac{\epsilon}{1 + \dfrac{|\epsilon|}{c}} \tag{27}$$

and as long as the residuals are small and c is small, the gradient will also be small. In both cases, g will be small, and from Eq. 26 we have

$$\|\gamma\| \leq K \|g\|, \tag{28}$$

where $K = \text{cond}(Y^T A)\|Y\|$, then $\|\gamma\| \approx 0$, meaning that all components of g will be small. The Hessian of the Lagrange function is given by

$$\nabla^2 L = \nabla^2 \phi + \sum_j \gamma_j H_j, \tag{29}$$

where $H_j = \nabla^2 f_j$. If the multipliers γ_j are small, then the Hessian of the Lagrangian can be approximated by the second derivatives of the objective function:

$$\nabla^2 L \approx \nabla^2 \phi. \tag{30}$$

Approximation 30 is called the Gauss–Newton (GN) approximation and it will be valid as long the residuals are small. Another case when the GN approximation will be valid is when the curvature of the constraints $f(x)$ is small. To see this, recall expression 29. If the curvature is small, then $\|H_j\| \approx 0$ and we get expression 30. However, for both the existence of a unique global solution to problem 24 and for the global convergence properties of the optimization algorithm, the reduced Hessian of the Lagrangian needs to be positive definite at the solution, and its approximations also need to be positive definite at every iteration. The reduced Hessian B is given by

$$B = Z^T \nabla^2 L Z, \tag{31}$$

where Z is some matrix whose columns span the null space of A^T. When using the GN approximation, positive definiteness of B will be guaranteed as long as the objective function is strictly convex.

For the log posterior of the contaminated normal function, the first derivative of the objective function will be approximated by and proportional to ϵ for small values of the residual ϵ, whereas for large ϵ it will be approximated by ϵ/b^2. If b is sufficiently large, then the second derivative will change signs as ϵ increases. Therefore this objective function is not convex and problem 24 may have many local minima even if the constraints are linear, and the GN approximation may not converge even to a local solution. On the other hand, the second derivative of the fair function 22 is given by

$$\frac{\partial^2 \rho(\epsilon, c)}{\partial \epsilon^2} = \frac{1}{(1 + |\epsilon|/c)^2}. \tag{32}$$

Note that this is a convex function with its curvature increasing with decreasing values of c. Here an estimation problem with constraints that have small curvature will always have a unique global solution and the GN approximation will be guaranteed to converge. Moreover, since the nonlinear data-reconciliation problem is frequently solved with SQP, the GN approximation leads to significant improvements in the convergence rate (Albuquerque and Biegler, 1996; Tjoa and Biegler, 1991). Unlike the contaminated normal objective, this is further improved by the fair function since it always has positive curvature. Therefore, from the optimization point of view, the fair function has more advantages than the nonconvex Bayesian approach.

Variable Classification

Variable classification is an important tool for designing instrumentation schemes and for providing more insight on how the measurements relate to the physical model being used. Gross-error detection relates to variable classification in the sense that special precautions have to be taken when dealing with nonredundant variables. Otherwise the reconciled solution will be nonunique and the optimization algorithm may fail. This is explored and illustrated further in the next section. In this section we first linearize and simplify the dynamic nonlinear model. We then apply the properties of observability and redundancy and derive the tools necessary to classify a variable.

Linearizing the DAE model

The DAE constraints in problem 1 are approximated by the general IRK formula:

$$z_{i+1} = z_i + \alpha z_i' \tag{33}$$

$$0 = F_i(z_i, z_i', u_i, \theta) \qquad i = 1, \ldots, N, \tag{34}$$

where z_i' are the stage derivatives. We solve for problem 1 with Eqs. 33 and 34 replacing the DAE. At the optimal point we linearize Eq. 34:

$$0 = A_i z_i - \phi_i z_i' + B_i u_i + \left(\frac{\partial F_i}{\partial \theta}\right)\theta + (\cdots) \tag{35}$$

where $A_i = (\partial F_i/\partial z_i)$, $B_i = (\partial F_i/\partial u_i)$, and $\phi_i = -(\partial F_i/\partial z_i')$, and $(\cdots)$ are the constant terms arising from Taylor's first-order expansion. We solve for z_i' using Eq. 35 inverting ϕ_i and replacing in Eq. 33. After some manipulation, we get:

$$z_{i+1} = A_i' z_i + B_i' u_i + C_i' \theta + (\cdots), \tag{36}$$

where $A_i' = I + \alpha\phi_i^{-1} A_i$, $B_i' = \alpha\phi_i^{-1} B_i$, $C_i = \alpha\phi_i^{-1}(\partial F_i/\partial\theta)$. We now group the state variables z_i and the input variables u_i into $x_i^T = (z_i^T, u_i^T)^T$. Equation 36 will lead to

$$[I \quad 0] x_{i+1} = [A_i' \quad B_i'] x_i + C_i' \theta + (\cdots). \tag{37}$$

Partitioning x_i into the measured variables x_i^m and unmeasured variables x_i^u, such that $x_i^T = [(x_i^m)^T, (x_i^u)^T]^T$ Eq. 37 will lead to

$$E_m x_{i+1}^m + E_u x_{i+1}^u = F_i^m x_i^m + F_i^u x_i^u + C_i' \theta + (\cdots), \tag{38}$$

where E_m and F_i^m are the matrices obtained by grouping the columns of $[I \quad 0]$ and $[A_i' \quad B_i']$ corresponding to the measured variables, respectively, and likewise E_u and F_i^u for the unmeasured variables. We do the same grouping and splitting for the initial and final conditions of Problem 1:

$$R_I E_m x_1^m + R_I E_u x_1^u = z_I \tag{39}$$

$$R_T E_m x_{N+1}^m + R_T E_u x_{N+1}^u = z_T. \tag{40}$$

Combining Eqs. 38, 39, and 40, we get

$$T^u x_u + T^m x_m = (\cdots), \tag{41}$$

where

$$x_u = \begin{bmatrix} x_1^u \\ \vdots \\ x_{N+1}^u \\ \theta \end{bmatrix}, \qquad x_m = \begin{bmatrix} x_1^m \\ \vdots \\ x_{N+1}^m \end{bmatrix}$$

$$T^u = \begin{bmatrix} (R_I E_u) & & & 0 \\ F_1^u & -E_u & & C_1' \\ & \ddots & & \\ & F_N^u & -E_u & C_N' \\ & & (R_T E_u) & 0 \end{bmatrix},$$

$$T^m = \begin{bmatrix} (R_I E_m) & & \\ F_1^m & -E_m & \\ & \ddots & \\ & F_N^m & -E_m \\ & & (R_T E_m) \end{bmatrix}.$$

Observability

Observability is closely related to identifiability (Seber and Wild, 1989). Here, given a nonlinear model $E[y] = f(t, \theta)$, where y are the response (measured) variables, t are the explanatory variables, the parameters θ will be unidentifiable when there exists θ_1 and θ_2 such that $f(t, \theta_1) = f(t, \theta_2)$. The concept of observability is very similar. An unmeasured variable is defined as unobservable if it cannot be uniquely determined through the measured variables. If a variable or a parameter is unobservable, then its value cannot be inferred from the measurements and the solution of the regression problem will be nonunique in these variables, and meaningless.

The application of observability results for steady-state systems (Crowe, 1989) to the dynamic case is straightforward. Consider Eq. 41, and assume a nonzero change Δx_u in the unmeasured variables x_u. Then $T^u \Delta x_u = 0$ for the perturbation to be feasible. Then $x_q \in x_u$ will be observable iff:

$$T^u \Delta x_u = 0 \Rightarrow \Delta x_q = 0. \tag{42}$$

An easy way to determine this is to decompose the matrix T^u into sparse LU factors

$$PT^u Q = L \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix},$$

where U_1 is an upper triangular nonsingular matrix of rank r, and P and Q are permutation matrices. Note that T^u may be full row rank. In this case, the zero entries in the upper triangular matrix would not exist. T^u could also be full column rank; then U_2 would not exist and there would not be any unobservable variables. In the more general case, the left-hand side of Eq. 42 will become

$$L \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix} Q^{-1} \Delta x_u = 0. \tag{43}$$

Rearranging vector Δx_u into

$$Q^{-1} \Delta x_u = \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix}$$

and simplifying Eq. 43, we get

$$\Delta x_1 = -U_1^{-1} U_2 \Delta x_2. \tag{44}$$

Because there will be a unique Δx_1 for any value of Δx_2 such that the change will be feasible and unobservable, then all the variables associated with Δx_2 will be unobservable. Furthermore, a variable associated with Δx_1 will be unobservable if the corresponding row of $-U_1^{-1} U_2$ is a nonzero vector, since it will always be possible to find a Δx_2 such that $\Delta x_1 \neq 0$. Observability requires a zero row of $U_1^{-1} U_2$. This approach is similar to the one developed by (Crowe, 1989) and (Swartz, 1989), except that we are now using it on dynamic systems with a sparse LU decomposition.

Redundancy

A variable is defined as nonredundant if deletion of its measurements will make this variable unobservable. In other

words, in order to be able to estimate the nonredundant variable, its measurements are absolutely necessary since this variable is not related to other measured variables through the model. For the simultaneous gross-error detection and data reconciliation, nonredundant variables play a primary role, since their detection as gross errors leads to unobservable variables and nonunique solution. Consequently, classification of these variables is crucial to the optimization strategy in the previous section.

Consider a local feasible change in x_u and x_m:

$$T^u \Delta x_u + T^m \Delta x_m = 0. \tag{45}$$

We now try to eliminate the unmeasured variables from Eq. 45. As was seen in the previous section, matrix T^u can be decomposed into LU factors:

$$PT^u Q = L \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

A matrix Z orthogonal to $(PT^u)^T$ would be given by

$$Z^T = [0 \quad I] L^{-1}. \tag{46}$$

Premultiplying Eq. 45 by the permutation matrix P and then by Z^T, the unmeasured variables would be eliminated and we would get

$$(Z^T P T^m) \Delta x_m = 0. \tag{47}$$

Suppose we have some measured variable x^k that is being measured across at least some of the $N+1$ data sets $x^k = (x_1^k, \ldots, x_{N+1}^k)^T \in x_m$ (e.g., temperature at the exit of a CSTR). Variable x^k will be nonredundant if by deletion of its associated measurements it becomes unobservable. We then apply the unobservability test to

$$(Z^T P T^m)_k \Delta x^k = 0, \tag{48}$$

where $(Z^T P T^m)_k$ are the columns of $(Z^T P T^m)$ corresponding to variable x^k. Previous studies (Crowe, 1989; Swartz, 1989) required the coefficient matrix in Eq. 47 to have a zero column for the associated variable to be nonredundant, and thus nonunique and unobservable if deleted. This condition is sufficient but not necessary for nonuniqueness in dynamic systems. Here a state or control variable is measured at several time instants; this also occurs in steady-state systems with multiple data sets. In both cases, the column associated with a particular measured variable (column of T^m) does not need to be zero. Instead the linear combination of all the columns associated with this variable (across all data sets or time instants) must lie in the null space of the coefficient matrix of the unmeasured variables (T^u). We further note that redundancy of a series of measurements requires that the columns of $(Z^T P T^m)_k$ span the entire space so that $\Delta x^k = 0$ is the only solution to Eq. 48 and a unique solution for x^k can be computed without the whole set of measurements.

In other words, the individual measurement for $z(t_i)$ may be redundant, even when $z(t) = [z(t_1), \ldots, z(t_{N+1})]$ is nonredundant. Here we are interested in the properties of variable $z(t)$, which upon discretization are equivalent to the *collective properties* of variables $(z_1, \ldots, z_{N+1})$. We term this concept *collective redundancy* for dynamic systems.

As an example consider the following process with only one measured variable z and one parameter θ, where the constraint is $z_{i+1} = z_i + \theta$. Obviously z is nonredundant, since there are no other measurements taken. If data are collected across data sets $i = 1, 2, 3$, then the constraints for the regression problem will be

$$z_3 = z_2 + \theta \tag{49}$$

$$z_2 = z_1 + \theta. \tag{50}$$

Bringing Eqs. 49 and 50 into matricial form, we get

$$\underbrace{\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}}_{T^m} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \underbrace{\begin{bmatrix} -1 \\ -1 \end{bmatrix}}_{T^u} \theta = 0. \tag{51}$$

A matrix Z orthogonal to T^u would be $Z^T = [1 \quad -1]$. Premultiplying Eq. 51 by Z^T, we get

$$\underbrace{[-1 \quad 2 \quad -1]}_{Z^T T^m} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = 0. \tag{52}$$

There are no zero columns in $Z^T T^m$; however, it is clear that (z_1, z_2, z_3) are nonunique when they lie in the null space of $[-1 \quad 2 \quad -1]$. Our test described previously would indicate (z_1, z_2, z_3) to be *collectively nonredundant*, although individual tests for z_1, z_2, or z_3 would classify them as redundant.

Now suppose that variables z_1 and z_2 were fixed to constant values through the addition of the following constraints:

$$z_1 = a \tag{53}$$

$$z_2 = b. \tag{54}$$

Obviously z_1, z_2, and z_3 are no longer collectively nonredundant. The coefficient matrices in Eq. 51 will be:

$$T^m = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$T^u = \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

and a matrix orthogonal to T^u would be:

$$Z = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{55}$$

Premultiplying Z^T by T^m, we get

$$Z^T T^m = \begin{bmatrix} -1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{56}$$

which is nonsingular. Therefore, in this case the test confirms that z_1, z_2, and z_3 are *collectively redundant.*

Characteristics of Robust Statistics

In the previous sections we described and compared both the Bayesian and the robust approaches for simultaneous gross-error detection. Since it models the gross errors, the Bayesian estimator deemphasizes outliers and detects them, by comparing the gross error and the random-error terms. On the other hand, robust estimators deemphasize outliers, as they are constructed to be resistant to deviations from the ideal error structures. However, a robust estimator provides no direct inferences to detect the outliers. First we describe a few techniques based on exploratory statistics that can be used to draw inferences based on the residuals produced by the robust regression. In the third section we saw that a careful analysis of the problem through variable classification can avoid numerical problems due to nonunique solutions of the regression problem. Following this subsection, we explore this topic a bit further and explain the numerical and statistical difficulties associated with performing gross-error detection on nonredundant variables and what can be done about it.

Identifying outliers and distributions: exploratory statistics

Once problem 1 is solved using the fair function or other robust estimator as the objective function, we end up with an estimate of the residuals $\{\epsilon_1, \ldots, \epsilon_N\}$. Since the robust estimators are resistant to outliers and to other anomalies, we assume that this sample will be representative of the measurement noise so that we can use batch exploratory methods to study them. Useful techniques for the distribution of the measurement noise are stem-and-leaf plots, dotplots, and probability plots. Stem-and-leaf and dotplots are essentially histograms, but with a different visual display. In dotplots, sample points within the same magnitude are plotted as points that pile up. In probability plots, the sample ordered statistics (residuals ordered in increasing values) are displayed vs. the order statistics (expected value of the ith observation in a sample of size N) of some distribution. If the sample is drawn from, say, a normal distribution, then it should be linear with the normal order statistics, the slope, and the intercept given by the sample standard deviation and mean. These plots not only are useful to give insight into the distribution of the measurement errors but are also useful to reveal outliers visually. A very simple and useful technique is the boxplot where the center of the box is the median and the sides are the quartiles. The outliers are spotted by computing order statistics (median and quartiles) and their distances from these. We start by computing the interquartile-range d_F:

$$d_F = F_u - F_l, \tag{57}$$

where F_u and F_l are the third and first quartiles, respectively. The outlier cutoffs are defined as $F_l - \alpha d_F$ and $F_u + \alpha d_F$, where α is usually set to 1/3. Measurements outside the cutoffs are considered outliers. A good description of these methods is given in Tukey et al. (1983). Once the outliers have been identified, this identification can always be verified by dropping the outliers and resolving a least-squares problem.

All these exploratory statistical tools are available in easy-to-use packages such as MINITAB (Ryan et al., 1985) or xlispstat (Tierney, 1990). In fact, all the quantile plots, dotplots, and boxplots associated with the examples below were produced using these packages.

Nonredundant variables: other robust estimators

In the previous section we discussed variable classification and we concluded that if a variable is unobservable, then it will not have a unique solution. Likewise, if a variable is nonredundant, then if the measurements associated with it are detected as outliers in the optimization, this variable will become unobservable and therefore nonunique. This leads to ill-behaved optimization problems where the solution will be difficult to obtain. Although these problems can be stabilized by using trust-region-based approaches (Dennis and Schnabel, 1983; Gopal and Biegler, 1995), the solution in these variables will be statistically meaningless as these nonredundant variables are not related to other measured variables through the model; their only sources of information are their own measurements. When using Bayesian methods or M estimators, these measurements are essentially ignored if they seem to be outliers, and this may happen even early in the optimization run. In this case a nonredundant variable will become unobservable and this can lead to convergence failures far from the solution.

Since the estimates of a nonredundant variable are totally dependent on their measurements, the contaminated normal or robust formulations cannot be used on these variables, as gross errors will make them unobservable and nonunique. Instead, robust estimates are needed based on replicate independent sensors measuring the same variable at the same time. With replication, the gross-error techniques described earlier can be used with less difficulty.

Moreover, although we focus mainly on M estimators in this article, there are other robust estimators that are more useful with replication when performing gross-error detection with nonredundant variables. L estimators, for instance, are linear combinations of order statistics. The simplest ones are the median for location and the median absolute deviation (MAD) for scale. Other statistics, such as the trimean, involve linear combinations of the quartiles. Although these statistics are very easy to compute, they are not differentiable. Another popular class of statistics are the R estimators, which are based on rank tests. Although they are more difficult to compute, they seem to have very good robustness and asymptotic properties. For more on these methods, refer to (Huber, 1980; Rey, 1983; Tukey et al., 1983; Rousseeuw and Leroy, 1987).

To deal with nonredundant variables, one can use an M estimator on replicated measurements, or one can preprocess the data for each time instant on this variable with L estimators, and feed the result to a least-squares estimator. This last approach seems more sensible since L estimators are very

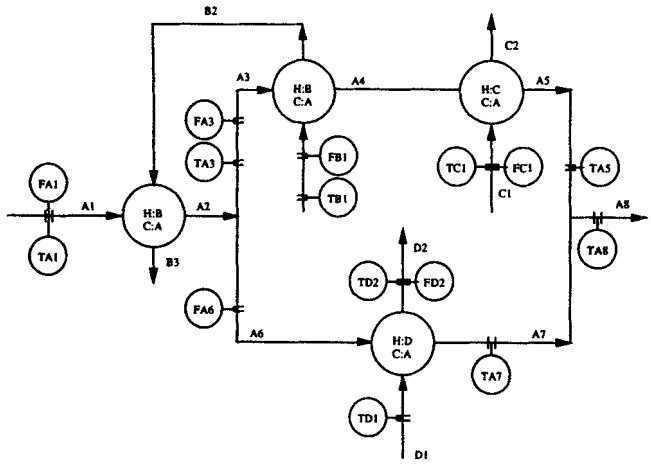

Figure 1. Heat exchanger network.

easy to use, and are extremely robust. If, for instance, we are measuring a nonredundant temperature, then we could use three separate thermocouples. For each time sample we would calculate the median (an L estimator), and this value would be given to the regression problem, where the objective function for this variable would be least squares. As shown in Examples 2 and 3, this robust statistic is highly resistant to outliers on nonredundant variables.

Finally, a simplification of this treatment occurs when the nonredundant variable is known to be constant with time, say, an input or control variable. Here we regard the values of this variable over time $\{u_1, \ldots, u_N\}$ as a sample from a statistical distribution and the batch exploratory techniques can be used to spot outliers. If an individual measurement u_i is detected as an outlier, then this measurement can be replaced by the median of the time-series samples. Moreover, if this input variable is a known function of time, then its measurements can be subtracted from the known function and the same approach can be applied.

Examples

In this section we present three examples. In the first example, we demonstrate the effectiveness of the simultaneous gross-error approach by comparing the results of both the Bayesian approach and the robust estimator with serial gross-error detection tests. Second, in the *tanks* problem we compare the fair function with the Bayesian approach, and we show how some exploratory tools can be used to spot the outliers and give some insight about the distributional structure of the measurement errors. In the final *hydrolysis* example, we show how gross-error detection can be performed with nonredundant data and also how to use both M and L estimators.

Heat exchanger network

Here we consider the steady-state heat-exchanger problem presented by Swartz (1989), where the gross-error detection

Table 1. Heat Exchanger Network: Regression Results

Variable Tag Name	Measurement	Std. Dev.	Swartz Run 1	Swartz Run 2	Contaminated Normal	Fair Function
FA1	1,000.00	20.00	963.63	969.12	968.81	954.25
TA1	466.63	0.75	466.33	466.33	466.33	466.33
FA2	—	—	963.63	969.12	968.81	954.25
TA2	—	—	481.91	481.77	481.78	481.79
FA3	401.70	8.03	407.86	406.68	406.66	401.95
TA3	481.78	0.75	481.81	481.77	481.78	481.79
FA4	—	—	407.86	406.68	406.66	401.95
TA4	530.09	0.75	530.09	530.09	530.09	530.09
FA5	—	—	407.86	406.68	406.66	401.95
TA5	616.31	0.75	615.51	616.31	616.27	616.29
FA6	552.70	11.05	555.77	562.44	562.15	552.30
TA6	—	—	481.91	481.77	481.78	481.79
FA7	—	—	555.77	562.44	562.15	552.30
TA7	619.00	0.75	617.76	613.94	614.09	618.80
FA8	—	—	963.63	969.12	968.81	954.25
TA8	614.92	0.75	616.81	614.93	615.01	617.74
FB1	253.20	5.06	253.20	253.20	253.20	253.20
TB1	618.11	0.75	618.11	618.11	618.11	618.11
FB2	—	—	253.20	253.20	253.20	253.20
TB2	—	—	543.90	543.91	543.85	544.75
FB3	—	—	253.20	253.20	253.20	253.20
TB3	—	—	486.51	486.71	486.59	488.37
FC1	308.10	6.16	308.10	308.10	308.10	308.10
TC1	694.99	0.75	694.99	694.99	694.99	694.99
FC2	—	—	308.10	308.10	308.10	308.10
TC2	—	—	594.80	594.10	594.15	595.34
FD1	—	—	689.42	679.72	680.09	693.35
TD1	667.84	0.75	668.02	667.83	667.84	667.86
FD2	680.10	13.60	689.42	679.72	680.09	693.35
TD2	558.34	0.75	558.17	558.35	558.34	558.32

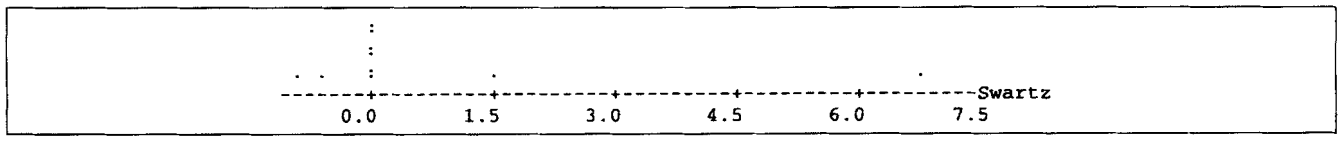

Figure 2. Dotplot for normalized residuals from Swartz regression.

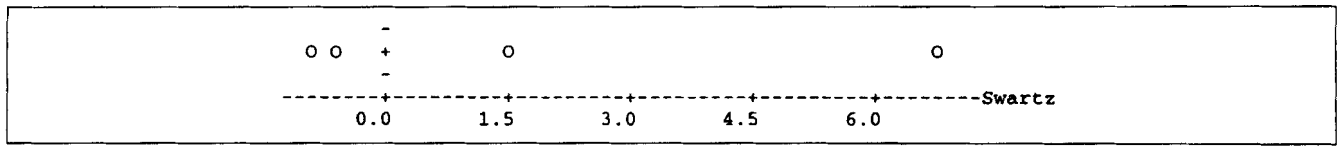

Figure 3. Boxplot for residuals from Swartz regression; median marked as +, probable outliers as ○.

was performed sequentially. A comparison between the serial approach and the simultaneous approach was done by Tjoa and Biegler (1991), where the contaminated normal distribution function 10 was used successfully to reconcile the data and detect the outliers. A simplified flowsheet is presented in Figure 1. In this problem the nonredundant variables are TA1, TA4, FB1, TB1, FC1, and TC1.

Table 1 shows the measurements, the standard deviations associated with them, and the estimates of the regressions performed in Swartz (1989), where a least-squares objective function was used on all measurements (run 1) and a measurement test was used to detect outliers. In this test, the measurement from variable TA7 was identified as a gross error. This measurement was then deleted and a new regression with the least squares was performed (run 2). No gross errors were detected this time. Table 1 also shows the results obtained in Tjoa and Biegler (1991) when the contaminated normal objective function is used with a probability of gross-error occurrence (η) equal to 0.05 and with a ratio of standard deviations (b) equal to 10. In this regression, the measurement of variable TA7 is identified as a gross error and the estimates are very close to those obtained in Swartz (1989) (run 2). The results from these two regressions are essentially the same. Finally, the same table shows the estimates of the regression when the fair function, Eq. 22, is used, tuned for an asymptotic relative efficiency of 70% for the normal distribution ($c = 0.04409$). Qualitatively similar results were also obtained with a 95% efficiency, but here we consider a lower efficiency for added robustness. This ensures that the difference between this estimator and the other methods will be more visible and better appreciated.

In all regressions the estimates for the nonredundant variables were the same as the measurements, which is not surprising in light of the discussion in the previous section. For the nonredundant variables, the results from the fair function were different from the ones obtained in Swartz (1989) and Tjoa and Biegler (1991). In order to explain the difference we plotted the residuals for the redundant variables in dotplots and boxplots (Figures 2–7). The dotplots and boxplots for the regressions with the contaminated normal and the serial approaches (Figures 2–5) clearly confirm the choice of TA7 as the source of outliers, as its residual appears as an extreme point. However, the dotplot and the boxplot for the residuals when using the fair function (Figures 6 and 7) indicate FA1 and TA8 as the source of gross errors that correspond to the extreme points in these plots. In Figure 8 we produce dotplots for the residuals from the three regressions with the outliers deleted. Note that the residuals for the regression with the fair function seem to be more compressed around the origin, which would indicate that the fair function picked a better choice of outliers than the other methods.

Two connected tanks

In this example, which we treated previously (Albuquerque and Biegler, 1996), we have two tanks connected by a valve. The measured variables are the flows F_0, F_1, F_2, and the levels of liquid h_1, h_2. The parameters are the inverse of the

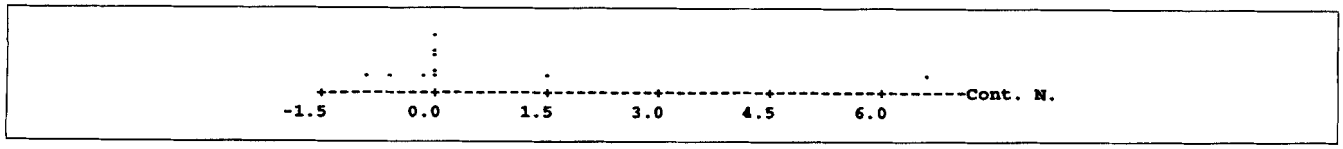

Figure 4. Dotplot for normalized residuals from regression with contaminated normal.

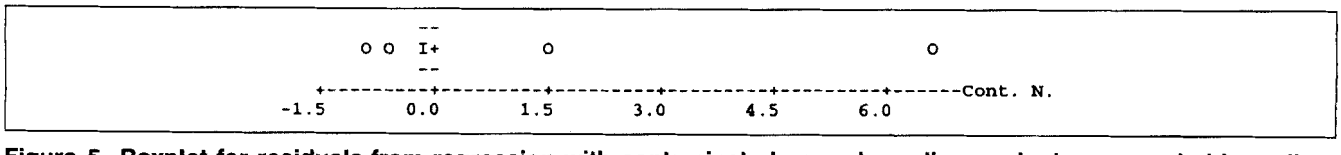

Figure 5. Boxplot for residuals from regression with contaminated normal; median marked as +, probable outliers as ○.

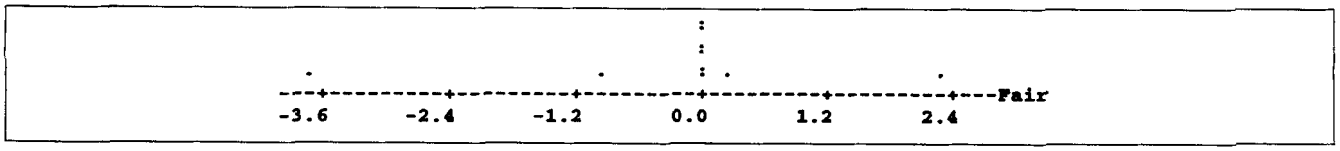

Figure 6. Dotplot for normalized residuals from regression with fair function.

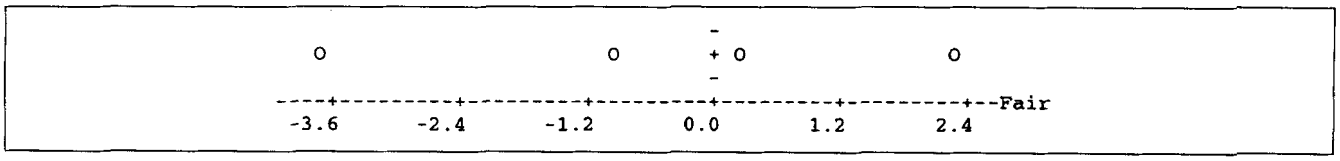

Figure 7. Boxplot for residuals from regression with fair function; median marked as +, probable outliers as ○.

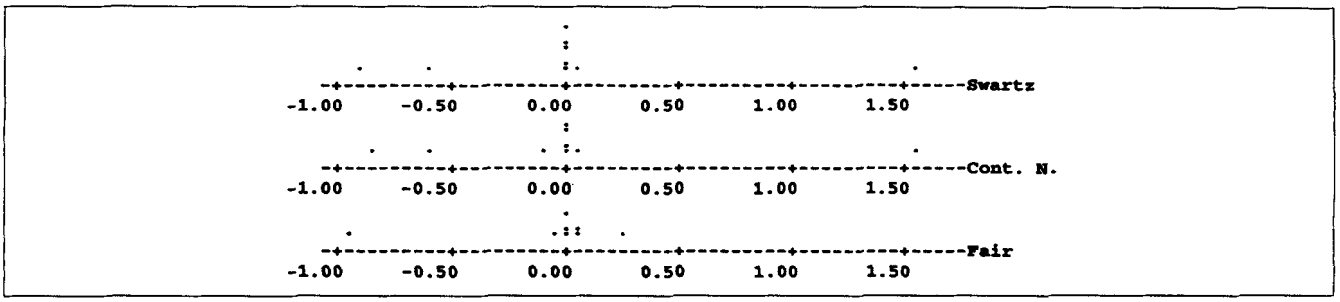

Figure 8. Dotplots for the normalized residuals from heat-exchanger network problem with the outliers removed.

cross-sectional areas $1/A_1$, $1/A_2$. The arrangement is shown in Figure 9. Using the techniques described in a previous section, no unobservable variables were detected. However, the input variable F_0 is nonredundant, as seen from the model for the two connected tanks:

$$A_1 h_1 = F_0 - F_1 \tag{58}$$

$$A_2 h_2 = F_1 - F_2 \tag{59}$$

$$h_1 = h_2 \tag{60}$$

$$F_2 = A_2 \sqrt{2gh_2}\,. \tag{61}$$

Both A_1 and F_0 can be estimated from Eq. 58 only. Unless F_0 is measured, then both F_0 and A_1 will be unobservable. In the first experiment, we simulated the data with normal noise along with large random shifts in the measured variables, except for F_0. These shifts simulate the gross errors. We performed the regression with our SQP decomposition strategies (Albuquerque and Biegler, 1996) using the least-squares function, the contaminated normal (Tjoa and Biegler, 1991) and the fair function tuned for an asymptotic relative efficiency of 95% for the normal distribution ($c = 1.3998$; see Eq. 16).

However, on all three runs the least-squares estimate was used for F_0. This variable is nonredundant, and to simplify the presentation we assume it has no gross errors. Deviation from this assumption will be examined later. The true values and the data for all the measured variables are shown in Figure 10. Table 2 shows the estimated values for the parameters and their relative errors for the least squares, contaminated normal, and robust fair function, and also the number

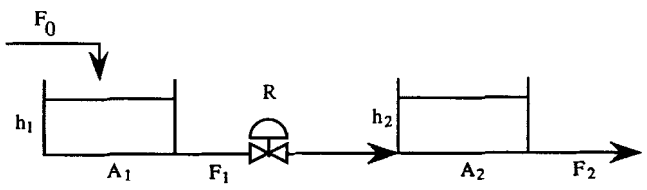

Figure 9. Two tanks connected by a valve.

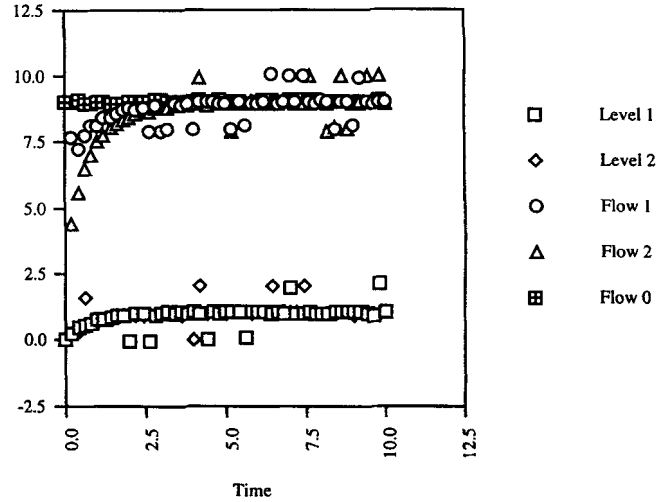

Figure 10. Data for tanks problem, random shifts as gross errors.

Table 2. Tanks Problem with Random Shifts as Gross Errors

Parameters	$1/A_1$	$1/A_2$	Iter.	CPU (seconds-VAX3200)
True value	0.500	0.500		
Least squares	0.698 (+40%)	0.503 (+0.6%)	8	35
Contaminated normal	0.490 (−2%)	0.500 (0%)	9	39
Fair function	0.501 (+0.2%)	0.501 (+0.2%)	8	33

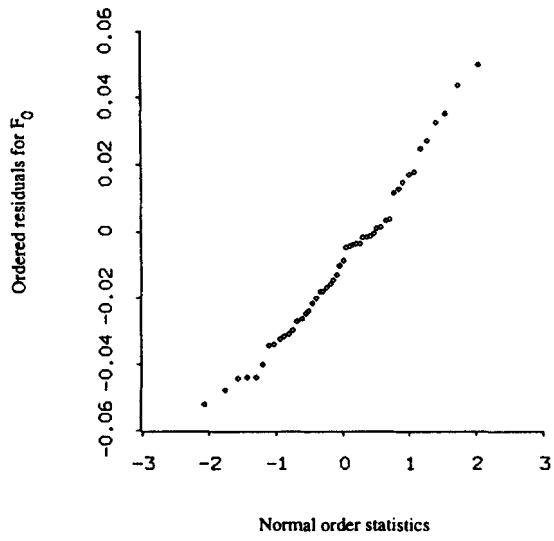

Figure 11. Normal probability plot for F_0.

of iterations required for convergence and the CPU time for a VAX-3200 workstation. The least squares showed a heavy bias, whereas both methods for handling the gross errors delivered excellent estimates. The contaminated normal method succeeded in detecting all the gross errors.

In Figures 11 to 15 we plot the residuals of F_0, F_1, F_2, h_1, h_2 obtained from the regression with the fair function on

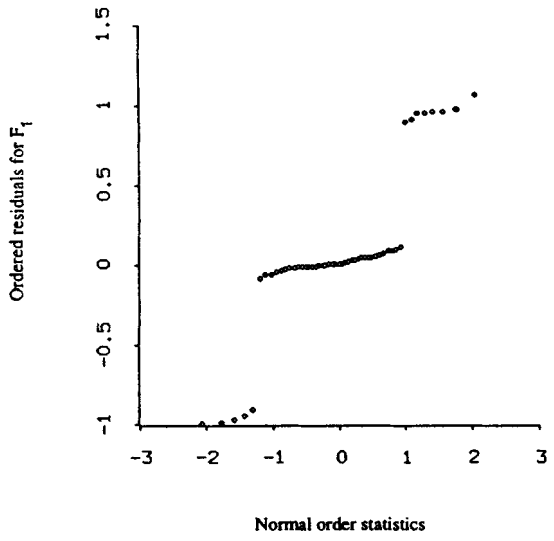

Figure 12. Normal probability plot for F_1.

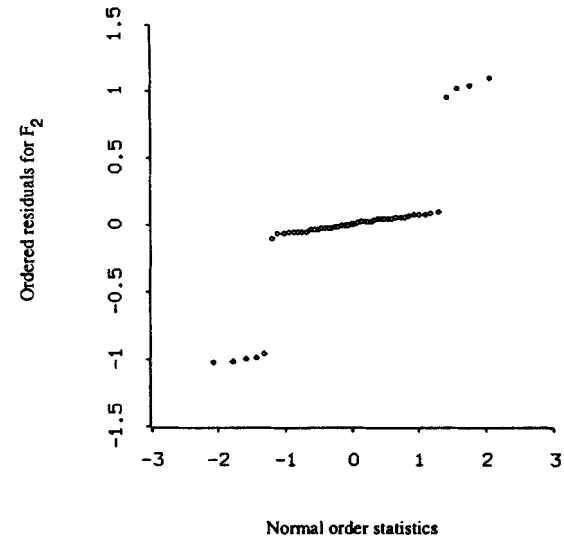

Figure 13. Normal probability plot for F_2.

normal quantile plots. These are probability plots where the residuals are plotted against the order statistics for the normal distribution (see the previous section). In these plots (Rawlings, 1988), linearity indicates normality. The outliers show up as distinct points in the extremes of the graphics. The bulk of the data are nearly normal (i.e., linear) in all variables confirming that the measurement error does follow a normal distribution. The CPU times for the three regressions differ very little in this case.

In another run, we simulated the data with normal random noise, but this time the measuring devices for the level in the first tank (h_1) and the flow measurements coming out of the second tank (F_2) were stuck at the second time period. Figures 16 and 17 show the true values, the data, and the reconciled values for these variables using the contaminated normal objective function and the fair function tuned for a relative asymptotic efficiency of 70% with the normal distribution ($c = 0.04409$). Both the fair function and the contaminated normal distribution captured the profile for F_2 quite

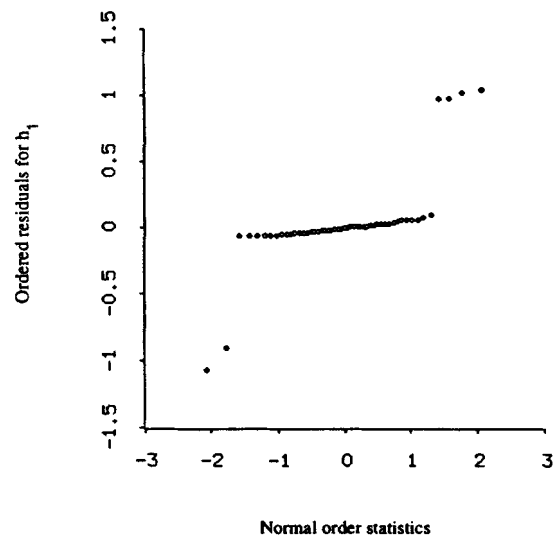

Figure 14. Normal probability plot for h_1.

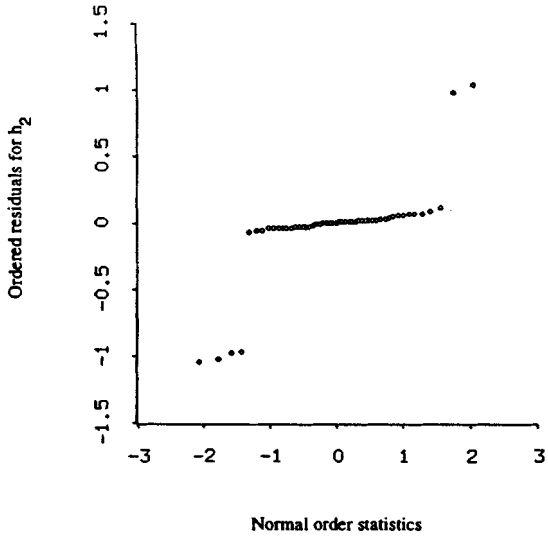

Figure 15. Normal probability plot for h_2.

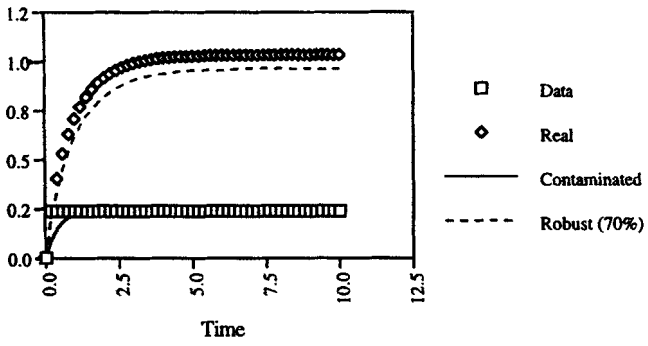

Figure 16. Tanks problem: stuck sensor on h_1.

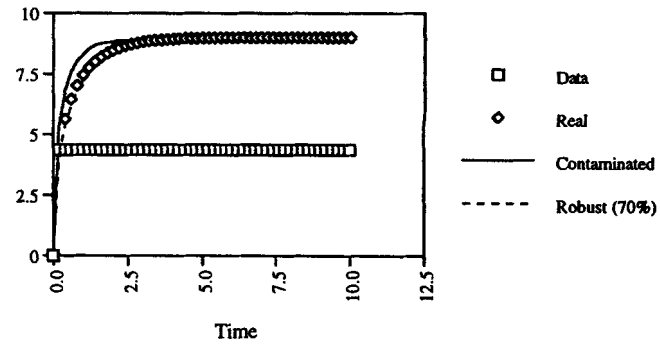

Figure 17. Tanks problem: stuck sensor on F_2.

Table 3. Tanks Problem with Stuck Sensors

Parameters	$1/A_1$	$1/A_2$	Iter.	CPU (seconds-VAX3200)
True value	0.500	0.500		
Least squares	0.25* (−50%)	0.571 (+14.2%)	10	43
Contaminated normal	0.25* (−50%)	0.25* (−50%)	19	83
Fair function, $E = 95\%$	0.25* (−50%)	0.456 (−8.8%)	15	66
Fair function, $E = 80\%$	0.314 (−37.2%)	0.476 (−4.8%)	16	70
Fair function, $E = 70\%$	0.337 (−32.6%)	0.482 (−1.8%)	24	105

*Bounds.

well (Figure 17); however, the contaminated normal failed to estimate the true values of h_1 and followed the errors instead, whereas the fair function did a much better job at ignoring the errors and estimating the true values (Figure 16).

In Table 3 we display the parameter estimates, number of iterations, and CPU time for the regression with the least squares, the contaminated normal distribution, and for the fair function tuned for relative asymptotic efficiencies with the normal distribution of 95, 80, and 70% ($c = 1.3998$, 0.17156, 0.04409). In all cases, the fair function did a better job than the contaminated normal and the least squares, with best results at the lowest efficiency. The contaminated normal performed worse than the least squares. This is not surprising since the gross error distribution is assumed normal and centered on zero residuals, which is clearly not the case, as the outlying measurements are heavily biased. Here the contaminated normal failed to detect the errors in h_1 (type I error), but flagged all measurements of h_2 as gross errors (type II error). By accepting the errors as valid data and discarding valid data as errors, the contaminated normal actually performed worse than the least squares, which does not make such distinctions. In Figures 18 to 22 we display dotplots of the residuals of the measured variables obtained with the fair function tuned at the lowest efficiency (70%). It can be seen that the residuals for F_0, F_1, and h_2 are concentrated around zero, whereas the residuals for F_2 and h_1 form a spike well away from zero, suggesting a heavy bias. Although the fair function required a computational effort 2.4 times larger than the least squares, it is still very cheap if we take into account the large number of gross errors detected in this example.

We now describe how to detect gross errors in F_0. In the previous examples we assumed that F_0 had no gross errors and we used a least-squared estimator since F_0 is a nonredundant variable. Use of either the fair function or the contaminated normal leads to poor convergence and performance of the optimization algorithm if F_0 is detected as an outlier. The measurements of this variable are not being replicated across the time instants. However, F_0 is also a control variable with a constant set point of 9.0, so gross-error detection can be performed without replication (see the

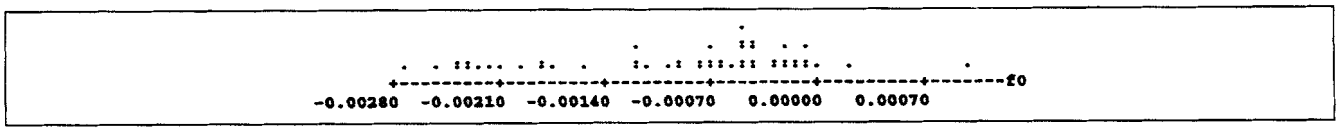

Figure 18. F_0 residuals for tanks problem with stuck sensors.

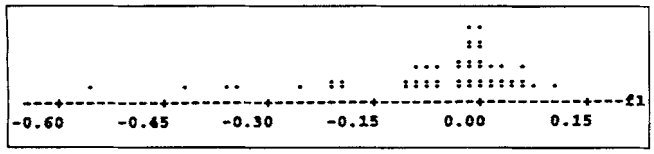

Figure 19. F_1 residuals for tanks problem with stuck sensors.

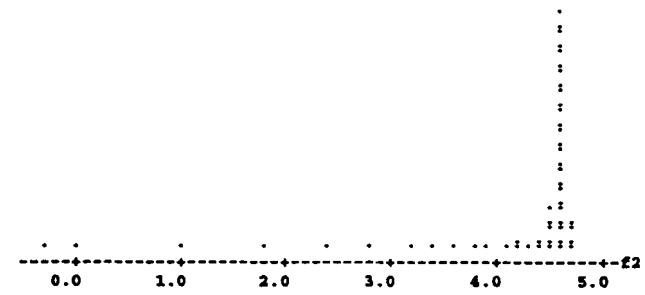

Figure 20. F_2 residuals for tanks problem with stuck sensors.

discussion in the previous section). We simulated the data for F_0 with random shifts as gross errors. The data and the errors are shown in Figure 23. A boxplot (Figure 24), with its outlier cutoffs computed as described earlier immediately spots the gross errors. These points can then be replaced by the median of the sample (9.02). Note that using least squares on this median is robust and resistant to outliers in F_0.

Acetic anhydride hydrolysis

In this problem, we have an isothermal viscous-flow tubular reactor with an acetic anhydride reaction and radial diffusion. The measurements are the bulk concentration at the end of the reactor and the unknown parameter is the kinetic constant (Cleland and Wilhelm, 1956). The model is the following:

$$(1-x^2)\frac{\partial u}{\partial t} = \beta\left(\frac{\partial^2 u}{\partial x^2} + \frac{1}{x}\frac{\partial u}{\partial x}\right) - ku \tag{62}$$

$$u(x,0) = 1 \tag{63}$$

$$\left.\frac{\partial u}{\partial x}\right|_{x=0,1} = 0, \tag{64}$$

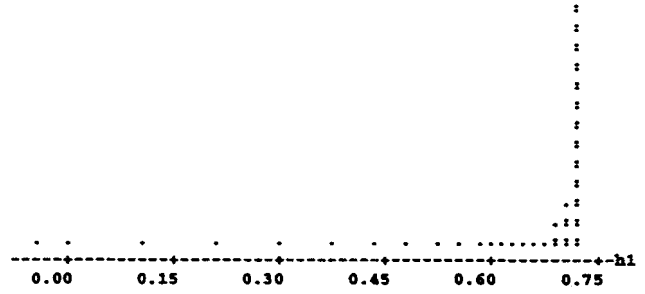

Figure 21. h_1 residuals for tanks problem with stuck sensors.

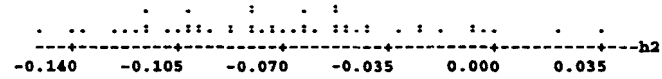

Figure 22. h_2 residuals for tanks problem with stuck sensors.

where $\beta = D/R^2$, $u = C/C_0$, $t = z/v_0$, $x = r/R$, and D is the diffusion coefficient; r is the radial direction starting from the center; R is the reactor's radius; C is the concentration at a given point; C_0 is the feed concentration; z is the axial coordinate; v_0 is the fluid velocity at the center of the reactor; and k is the kinetic constant to be estimated. Although this problem is in steady state, Eqs. 62, 63, and 64 form a parabolic partial differential equation that can be converted into an ODE system with initial conditions in t, using the method of weighted residuals (Ames, 1992). It can thus be solved efficiently by our decomposition techniques (Albuquerque and Biegler, 1996). More details on how to treat distributed systems will be left to a future article. The measured variable is the bulk concentration C_a at the length L of the reactor.

$$C_a = 4\int_0^1 x(1-x^2)u(x,L)\,dx. \tag{65}$$

These measurements are collectively nonredundant; therefore for robust gross-error techniques to be used, they must be replicated. We show how estimators other than M estimators can be used in a situation where the nonredundant data are replicated and therefore do not have to be taken at face value. Given several measurements of $(C_a^1, \ldots, C_a^m)$, we can

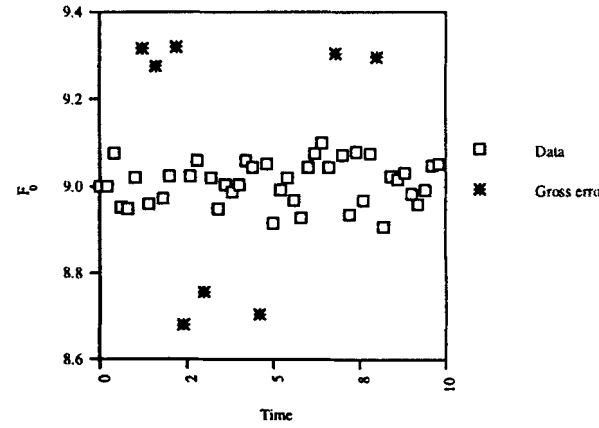

Figure 23. Tanks problem: random shifts as gross errors on F_0.

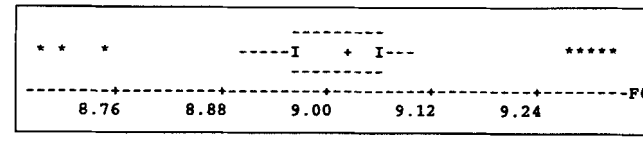

Figure 24. Tanks problem: boxplot for F_0 with random shifts as gross errors; data in Figure 23; median marked as +, quartiles as I, outliers as *.

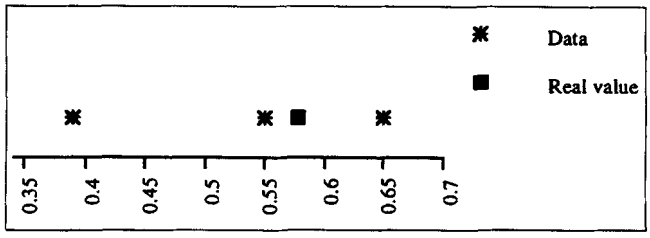

Figure 25. Hydrolysis problem, small bias in one data point.

Table 4. Hydrolysis Problem, Small Bias in One Data Point

Parameter	k	Iter.	CPU (seconds-VAX3200)
True value	0.2733		
Least squares	0.3216 (+17.7%)	15	20
Fair function, $E = 95\%$	0.3127 (+14.4%)	13	18
Fair function, $E = 70\%$	0.3010 (+10.1%)	19	25
Least squares with median	0.3009 (+10.1%)	14	18

use an M estimator, or we can preprocess the sample using L-estimators (e.g., the median of the replicates) and feed this estimate to a least-squares estimator. In the first run we used three points, two of them close to the real value of the exit concentration, and one somewhat farther away (Figure 25). We then used the least-squares function, the fair function tuned for 95% and 70% relative asymptotic efficiencies for the normal distribution ($c = 1.3998$, 0.04409) using the sample standard deviation as the scale measure, and the least squares preprocessed by the median (L estimator).

The estimates are compared in Table 4 along with the required number of iterations and CPU time on a VAX-3200 workstation. The fair function performed better than the least squares and as well as the preprocessed least squares when tuned for a low relative asymptotic efficiency. All regressions required comparable computational effort. In another run, we pulled the biased observation farther away from the concentration's true value (Figure 26). The parameter estimates for this case are shown in Table 5 along with the computational loads. Although the fair function still behaves very well when tuned at a 70% efficiency, it loses some accuracy in the estimate, whereas the preprocessed least squares totally ignores the added bias. In this last approach we are taking advantage of the high robustness of L estimators. Again, there is not much difference in the computational effort of the different regressions.

Conclusions

While data reconciliation is a commonly used tool for steady-state systems, its application to dynamic systems is still in its infancy. In this article, we explore several aspects of the dynamic data-reconciliation problem, including gross-error detection and variable classification. For this study we see that both aspects are strongly linked and have an important impact on the solution of the optimization problem and performance of the optimization algorithm.

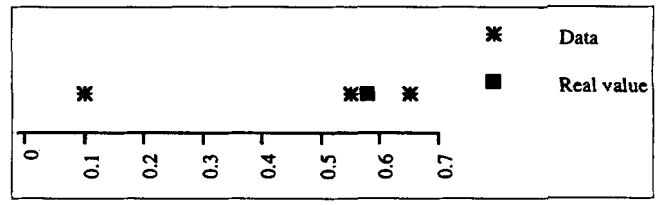

Figure 26. Hydrolysis problem, large bias in one data point.

Table 5. Hydrolysis Problem, Large Bias in One Data Point

Parameter	k	Iter.	CPU (seconds-VAX3200)
True value	0.2733		
Least squares	0.4375 (+60.1%)	14	18
Fair function, $E = 95\%$	0.3846 (+40.7%)	13	18
Fair function, $E = 70\%$	0.3017 (+10.4%)	16	22
Least squares with median	0.3009 (+10.1%)	14	18

For gross-error detection we embed and analyze robust estimators within our data-reconciliation formulation and compare this approach to a previously studied contaminated normal formulation. The robust approach has a number of advantages, including better numerical characteristics and less biased estimates. Moreover, as analyzed in this study this approach has the interesting property of yielding global solutions for nonlinear programs with low constraint curvature.

For data reconciliation, uniqueness of the optimal solution and the consequent convergence to this solution is closely tied to observability and redundancy of the measurements. This analysis has been revisited and developed with cheap LU factorizations, and, for dynamic systems, we introduce the new concept of collective redundancy, which is relevant when the time series disappears. This analysis is therefore useful not only for interpreting the reconciled values but also for successful performance of the optimization algorithm.

Finally, one feature of the robust approach is that no error structure is assumed from which inferences can be drawn. As a result, exploratory statistics must be used to aid in the interpretation of the results. Three examples are analyzed to illustrate all of these concepts and demonstrate the advantages of the robust approach for dynamic data reconciliation.

Acknowledgment

One of the authors (J. S. A.) was supported by a fellowship grant from Programa CIÊNCIA, JNICT, Lisbon, Portugal. We are also grateful to an anonymous reviewer for sharpening the concept of collective redundancy.

Literature Cited

Albuquerque, J. S., and L. T. Biegler, "Decomposition Algorithms for On-line Estimation with Nonlinear DAE Models," *Comput. Chem. Eng.*, in press (1996).

Ames, W. F., *Numerical Methods for Partial Differential Equations*, 3rd ed., Academic Press, San Diego (1992).

Basu, A., and K. K. Paliwal, "Robust M-Estimates and Generalized M-Estimates for Autoregressive Parameter Estimation," TENCON 89, 4th IEEE Region 10 Int. Conf., Bombay (1989).

Brenan, K. E., S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Elsevier, New York (1989).
Cleland, F. A., and R. H. Wilhelm, "Diffusion and Reaction in a Viscous-flow Tubular Reactor," *AIChE J.*, **2**(4), 489 (1956).
Crowe, C. M., "Observability and Redundancy of Process Data for Steady State Reconciliation," *Chem. Eng. Sci.*, **44**(12), 2909 (1989).
DeGroot, M. H., *Probability and Statistics*, 2nd ed., Addison-Wesley, Reading, MA (1986).
Dennis, J. E., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, Englewood Cliffs, NJ (1983).
Fariss, R. H., and V. H. Law, "An Efficient Computational Technique for Generalized Application of Maximum Likelihood to Improve Correlation of Experimental Data," *Comput. Chem. Eng.*, **3**, 95 (1979).
Gopal, V., and L. T. Biegler, "Nonsmooth Dynamic Simulation with Linear Programming Based Methods," *Comput. Chem. Eng.*, submitted (1995).
Huber, P. J., *Robust Statistics*, Wiley, New York (1980).
Jeffreys, H., "An Alternative to the Rejection of Observations," *Proc. Roy. Soc.*, **A137**, 78 (1932).
Johnston, L. P. M., and A. M. Kramer, "Maximum Likelihood Data Rectification: Steady State Systems," *AIChE J.*, **41**(11), 2415 (1995).
Kretsovalis, A., and R. S. H. Mah, "Observability and Redundancy Classification in Generalized Process Networks: I. Theorems," *Comput. Chem. Eng.*, **12**, 671 (1988).
Liebman, M. J., T. F. Edgar, and L. S. Lasdon, "Efficient Data Reconciliation and Estimation for Dynamic Processes using Nonlinear Programming Techniques," *Comput. Chem. Eng.*, **16**(10/11), 693 (1992).
Narasimhan, S., and R. S. H. Mah, "Generalized Likelihood Ratios for Gross Error Identification in Dynamic Processes," *AIChE J.*, **34**(8), 1321 (1988).
Rawlings, J. O., *Applied Regression Analysis. A Research Tool*, Wadsworth & Brooks, Pacific Grove, CA (1988).
Rey, W. J. J., *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag, Berlin/New York (1983).
Rousseeuw, P. J., and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York (1987).
Ryan, B. F., L. B. Joiner, and T. A. Ryan, *MINITAB Handbook*, 2nd ed., PWS-Kent, Boston (1985).
Seber, G. A. F., and C. J. Wild, *Nonlinear Regression*, Wiley, New York (1989).
Sistu, P. B., R. S. Gopinath, and B. W. Bequette, "Computation Issues in Nonlinear Predictive Control," *Comput. Chem. Eng.*, **17**(4), 361 (1993).
Stanley, G. M., and R. S. H. Mah, "Observability and Redundancy in Process Data Estimation," *Chem. Eng. Sci.*, **36**, 259 (1981).
Swartz, C. L. E., "Data Reconciliation for Generalized Flowsheet Applications," ACS Meeting, Dallas (1989).
Tamhane, A. C., C. Kao, and R. S. H. Mah, "Gross Error Detection in Serially Correlated Process Data. 2. Dynamic Systems," *Ind. Eng. Chem. Res.*, **31**, 254 (1992).
Tierney, L., *Lisp-Stat*, Wiley, New York (1990).
Tjoa, I. B., and L. T. Biegler, "Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems," *Comput. Chem. Eng.*, **15**(10), 679 (1991).
Tukey, J. W., D. C. Hoaglin, and F. Mosteller, *Understanding Robust and Exploratory Data Analysis*, Wiley, New York (1983).
Verdinelli, I., and L. Wasserman, "Bayesian Analysis of Outlier Problems using the Gibbs Sampler," *Stat. Comput.*, **1**, 105 (1991).

Manuscript received Aug. 14, 1995, and revision received Feb. 28, 1996.