Mass Balance Equilibration: A Robust Approach Using Contaminated Distribution

José Ragot, Mohammed Chadli, and Didier Maquin
Centre de Recherche en Automatique de Nancy, CNRS UMR 7039, Institut National Polytechnique de Lorraine, F-54 516 Vandoeuvre les Nancy Cedex, France

DOI 10.1002/aic.10412
Published online April 1, 2005 in Wiley InterScience (www.interscience.wiley.com).

Keywords: data reconciliation, contaminated distribution, process control

Introduction

The problem of obtaining reliable estimates of the state of a process is a fundamental objective in process supervision, given that these estimates are used to understand the process behavior. Measurements are collected to know, at each time, the behavior of the process and provide a way to verify whether its functioning has been defined and given by the user. For that purpose, broadly ranging techniques have been developed to perform what is currently known as *data reconciliation* (for which several book reviews have been written[1-4]): the underlying idea is to verify whether the measurements fulfill the model of the process and, if this is not the case, to analyze what are the noises affecting the measurements and finally to correct the measurements. Unfortunately, the measurements may be unknowingly corrupted by gross errors, the effects of which are added to those of the noise. As a result, the data reconciliation procedure can give rise to absurd results and the estimated variables are corrupted by this bias. Several schemes have been suggested to cope with the corruption of normal assumption of the errors,[1,5,6] by detecting a priori or a posteriori gross errors.

Methods to include bounds in process variables to improve gross error detection have been developed. [7] However, if bounds are not properly chosen, one major disadvantage of these methods is that they give rise to situations such that it may impossible to estimate all the variable using only a subset of the remaining free gross error measurements. There is also an important class of robust estimators whose influence function is bounded, thus allowing rejection of outliers.[8,9] Another approach is to take into account the nonideality of the measurement error distribution using an objective function constructed on contaminated error distribution.[8,10] This approach has been developed for data reconciliation[11,12] and has been tested on several applications.[6,13-15]

In this article, we restrict our analysis to processes described by linear and bilinear mass balance equations. However, despite this limitation, these models are in current use because they describe total mass and partial mass balances. In the following, we adopt and develop the use of contaminated data distribution for the data reconciliation problem. The second section provides a brief background of the data reconciliation problem and, in the third section, the proposed robust data reconciliation method is developed. It is thereafter illustrated through an academic example in the fourth section.

Data Reconciliation Background: The Linear Case

The classical general data reconciliation problem[2,16-18] deals with a weighted least-square minimization of the measurement adjustments subject to the model constraints. Indeed, for the sake of simplicity, the process model equations are taken as linear:

$$Ax = 0 \qquad A \in \mathbb{R}^{nv} \qquad x \in \mathbb{R}^{v} \tag{1}$$

where x is the state of the process and A is its so-called incidence matrix. Measurement gives the partial information $\tilde{x} \in \mathbb{R}^{p}$:

$$\tilde{x} = Hx + \varepsilon \qquad \tilde{x} \in \mathbb{R}^{p} \tag{2a}$$

$$\varepsilon \sim N(0, V) \qquad H \in \mathbb{R}^{pv} \tag{2b}$$

where the matrix H defines which variables are measured and where $\varepsilon \in \mathbb{R}^{p}$ is a vector of random errors, characterized by a variance matrix V and a normal probability density function (pdf):

Correspondence concerning this article should be addressed to J. Ragot at jragot@ensem.inpl-nancy.fr.

© 2005 American Institute of Chemical Engineers

$$p_x = \frac{1}{(2\pi)^{v/2}\sqrt{\det(V)}} \exp\left(-\frac{1}{2}\|Hx - \tilde{x}\|^2_{V^{-1}}\right) \quad (3)$$

The likelihood estimation is obtained by optimizing the Lagrange function:

$$\mathcal{L} = \frac{1}{(2\pi)^{v/2}\sqrt{\det(V)}} \exp\left(-\frac{1}{2}\|Hx - \tilde{x}\|^2_{V^{-1}}\right) + \lambda^T Ax \quad (4)$$

Assuming the system observability [$\text{rank}(A^T\ H^T) = n$], the well-known solution of this problem is[19]:

$$\hat{x} = [G - GA^T(AGA^T)^{-1}AG]H^T V^{-1}\tilde{x} \quad (5a)$$

$$G = (H^T V^{-1} H + A^T A)^{-1} \quad (5b)$$

Extensions for dynamic systems were also established (see, for example, Singhal and Seborg[12] and Soderstrom et al.[14]). In fact, the estimations obtained by this method are not always exploitable, the main drawback of which is the contamination of all estimated values by the outliers. For that reason robust estimators could be preferred, where *robustness* is the ability to ignore the contribution of extreme data such as gross errors. Robust statistics[8-10,20] treat the consequences of possible deviation from the statistical model, providing methods for protecting data reconciliation procedures against such deviations. In this field, one of the pioneer works[11] uses a method based on a contaminated Gaussian objective function instead of the classical least-square objective function.

Robust Data Validation: The Bilinear Case

We consider now the case of a process characterized by two types of variables: macroscopic variables (such as flow rates x) and microscopic variables (such as concentrations or particle sizes y). Thus, the process model (Eq. 1) is extended to

$$Ax = 0 \qquad A \in \mathbb{R}^{nv} \qquad x \in \mathbb{R}^v \quad (6a)$$

$$A(x \otimes y) = 0 \qquad y \in \mathbb{R}^v \quad (6b)$$

The operator $\otimes$ is used to perform the element by element product of two vectors and thus describes compactly bilinear equations. In this section, all the process variables are assumed to be measured; so $\tilde{x} \in \mathbb{R}^v$ and $\tilde{y} \in \mathbb{R}^v$.

If the measurements contain random outliers, then a single pdf described as in Eq. 3 cannot account for the high variance of the outliers. To overcome this problem let us assume that measurement noise is sampled from two pdfs, the normal one having a small variance representing regular noise and the abnormal one having a large variance representing outliers. To simplify the presentation, each measurement $\tilde{x}_i$ ($\tilde{y}_i$) is assumed to have the same normal $\sigma_{x,1}$ ($\sigma_{y,1}$) and abnormal $\sigma_{x,2}$ ($\sigma_{y,2}$) standard deviations. This hypothesis will be withdrawn later. Thus, for each observation $\tilde{x}_i$ and $\tilde{y}_i$, we define the following pdf:

$$p(\tilde{x}_i|x_i,\ \sigma_{x,j}) = \frac{1}{\sqrt{2\pi}\sigma_{x,j}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \tilde{x}_i}{\sigma_{x,j}}\right)^2\right] \qquad j = 1, 2,\ i = 1 \cdots v \quad (7a)$$

$$p(\tilde{y}_i|y_i,\ \sigma_{y,j}) = \frac{1}{\sqrt{2\pi}\sigma_{y,j}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \tilde{y}_i}{\sigma_{y,j}}\right)^2\right] \qquad j = 1, 2,\ i = 1 \cdots v \quad (7b)$$

In the following, we adopt the shortening notation $p_{x,j,i}$ and $p_{y,j,i}$, respectively, for $p(\tilde{x}_i \,|\, x_i,\ \sigma_{x,j})$ and $p(\tilde{y}_i \,|\, y_i,\ \sigma_{y,j})$ where indices i and j are, respectively, used to indicate the number of data and the number of the distribution. Then, the combination of these two pdfs (for each type of variable) is performed with the help of a weight w. Quantity $(1 - w)$ can be seen as an a priori probability of the occurrence of outliers:

$$p_{x,i} = w p_{x,1,i} + (1 - w) p_{x,2,i} \qquad i = 1 \cdots v \quad (8a)$$

$$p_{y,i} = w p_{y,1,i} + (1 - w) p_{y,2,i} \qquad i = 1 \cdots v \quad (8b)$$

Assuming independence of the measurements allows the global log-likelihood function to be defined as

$$\Phi = \log \prod_{i=1}^{v} p_{x,i} p_{y,i} \quad (9)$$

Let us now define the optimization problem consisting in estimating the process variables x and y. For that, consider the following Lagrange function:

$$\mathcal{L} = \Phi + \lambda^T Ax + \mu^T A(x \otimes y) \quad (10)$$

Mass balance constraints for total flow rate (Eq. 6a) and partial flow rate (Eq. 6b) are taken into account through the introduction of the parameters λ and μ. The stationarity conditions of $\mathcal{L}$, with respect to x, y, λ, and μ, are expressed by direct derivative (the estimations are now noted $\hat{x}$ and $\hat{y}$):

$$W_{\hat{x}}^{-1}(\hat{x} - \tilde{x}) + A^T\lambda + (A \otimes \hat{y})^T\mu = 0 \quad (11a)$$

$$W_{\hat{y}}^{-1}(\hat{y} - \tilde{y}) + (A \otimes \hat{x})^T\mu = 0 \quad (11b)$$

$$A\hat{x} = 0 \quad (11c)$$

$$A(\hat{x} \otimes \hat{y}) = 0 \quad (11d)$$

The weighting matrices $W_{\hat{x}}$ and $W_{\hat{y}}$ are defined by

$$W_{\hat{x}}^{-1} = \underset{i=1\cdots v}{\text{diag}}\left[\frac{w p_{\hat{x},1,i}}{\sigma_{x,1}^2} + \frac{(1-w) p_{\hat{x},2,i}}{\sigma_{x,2}^2} \over w p_{\hat{x},1,i} + (1-w) p_{\hat{x},2,i}\right] \quad (12a)$$

$$W_{\hat{y}}^{-1} = \underset{i=1\cdots v}{\text{diag}}\left[\frac{wp_{\hat{y},1,i} + \dfrac{(1-w)p_{\hat{y},2,i}}{\sigma_{y,1}^2}}{wp_{\hat{y},1,i} + (1-w)p_{\hat{y},2,i}}\right] \tag{12b}$$

where the notations $\underset{i=1\cdots v}{\text{diag}(a_i)}$ or $\text{diag}(a)$ stand for the operator that converts a v-dimensional vector a, whose entries are a_i, into a diagonal matrix. Notice that if each measurement $\tilde{x}_i$ ($\tilde{y}_i$) has a particular standard deviation, Eqs. 12a and 12b still hold by replacing the parameters $\sigma_{x,1}$ and $\sigma_{x,2}$ ($\sigma_{y,1}$ and $\sigma_{y,2}$) by $\sigma_{x,1,i}$ and $\sigma_{x,2,i}$ ($\sigma_{y,1,i}$ and $\sigma_{y,2,i}$). Using shortening notations $A_{\hat{x}} = A\ \text{diag}(\hat{x})$ and $A_{\hat{y}} = A\ \text{diag}(\hat{y})$, the system of Eqs. 11a–11d may be reformulated as an implicit system with respect to the unknowns $\hat{x}$ and $\hat{y}$

$$\hat{x} = [I - W_{\hat{x}}A^T(AW_{\hat{x}}A^T)^{-1}A][\tilde{x} - W_{\hat{x}}A_{\hat{y}}^T(A_{\hat{x}}W_{\hat{y}}A_{\hat{x}}^T)^{-1}A_{\hat{x}}\tilde{y}] \tag{13a}$$

$$\hat{y} = [I - W_{\hat{y}}A_{\hat{x}}^T(A_{\hat{x}}W_{\hat{y}}A_{\hat{x}}^T)^{-1}A_{\hat{x}}]\tilde{y} \tag{13b}$$

Equations 13a and 13b are clearly nonlinear with respect to the unknowns $\hat{x}$ and $\hat{y}$, the weight matrices $W_{\hat{x}}$ and $W_{\hat{y}}$, depending on the Eq. 8 pdf, which themselves depend on the $\hat{x}$ and $\hat{y}$ estimations. In fact, the solution of the implicit system of Eqs. 13a and 13b can be numerically obtained with standard solvers; for more efficiency, we suggest the following iterative scheme, which is well adapted to the specific bilinear structure of the equations.

Initialization Step. $k = 0$, $\hat{x}^{(k)} = \tilde{x}$, $\hat{y}^{(k)} = \tilde{y}$. Based on an a priori knowledge about the occurrence of gross errors, choose w. Adjust $\sigma_{x,1}$ and $\sigma_{y,1}$ from an a priori knowledge about the noise distribution, or take them proportional to the measurements. Adjust $\sigma_{x,2}$ and $\sigma_{y,2}$ from an a priori knowledge about the gross error distribution or take them proportional to the measurements and greater that $\sigma_{x,1}$ and $\sigma_{y,1}$.

Estimation Step. Compute the quantities

$$p_{\hat{x},j,i}^{(k)} = \frac{1}{\sqrt{2\pi}\sigma_{x,j}}\exp\left\{-\frac{1}{2}\left[\frac{\hat{x}_i^{(k)} - \tilde{x}_i}{\sigma_{x,j}}\right]^2\right\} \qquad j = 1, 2,\ i = 1\cdots v \tag{14a}$$

$$p_{\hat{y},j,i}^{(k)} = \frac{1}{\sqrt{2\pi}\sigma_{y,j}}\exp\left\{-\frac{1}{2}\left[\frac{\hat{y}_i^{(k)} - \tilde{y}_i}{\sigma_{y,j}}\right]^2\right\} \qquad j = 1, 2,\ i = 1\cdots v \tag{14b}$$

$$W_{\hat{x}}^{-1} = \underset{i=1\cdots v}{\text{diag}}\left[\frac{wp_{\hat{x},1,i}^{(k)} + \dfrac{(1-w)p_{\hat{x},2,i}^{(k)}}{\sigma_{x,1}^2}}{wp_{\hat{x},1,i}^{(k)} + (1-w)p_{\hat{x},2,i}^{(k)}}\right]$$

$$W_{\hat{y}}^{-1} = \underset{i=1\cdots v}{\text{diag}}\left[\frac{wp_{\hat{y},1,i}^{(k)} + \dfrac{(1-w)p_{\hat{y},2,i}^{(k)}}{\sigma_{y,1}^2}}{wp_{\hat{y},1,i}^{(k)} + (1-w)p_{\hat{y},2,i}^{(k)}}\right] \tag{14c}$$

$$A_{\hat{x}}^{(k)} = A\ \text{diag}[\hat{x}^{(k)}] \qquad A_{\hat{y}}^{(k)} = A\ \text{diag}[\hat{y}^{(k)}] \tag{14d}$$

Update the estimations

$$\hat{x}^{(k+1)} = \{I - W_{\hat{x}}^{(k)}A^T[AW_{\hat{x}}^{(k)}A^T]^{-1}A\} \times \{\tilde{x} - W_{\hat{x}}^{(k)}A_{\hat{y}}^{(k)T}[A_{\hat{x}}^{(k)}W_{\hat{y}}^{(k)}A_{\hat{x}}^{(k)T}]^{-1}A_{\hat{x}}^{(k)}\tilde{y}\} \tag{15a}$$

$$\hat{y}^{(k+1)} = \{I - W_{\hat{y}}^{(k)}A_{\hat{x}}^{(k)T}[A_{\hat{x}}^{(k)}W_{\hat{y}}^{(k)}A_{\hat{x}}^{(k)T}]^{-1}A_{\hat{x}}^{(k)}\}\tilde{y} \tag{15b}$$

Convergence Test Step. Compute an appropriate norm of the corrective terms: $\tau_x^{(k+1)} = \|\hat{x}^{(k+1)} - \tilde{x}\|$ and $\tau_y^{(k+1)} = \|\hat{y}^{(k+1)} - \tilde{y}\|$. If the variations $\tau_x^{(k+1)} - \tau_x^{(k)}$ and $\tau_y^{(k+1)} - \tau_y^{(k)}$ are less than a given threshold then stop, else $k = k + 1$ and go to step 2.

Extensions

Partial measurements

Let us consider the more realistic situation where only some variables are measured. For that purpose, two selection matrices H_x and H_y are introduced, thus allowing us to define which variables are measured:

$$\tilde{x} = H_x x + \varepsilon_x \tag{16a}$$

$$\tilde{y} = H_y y + \varepsilon_y \tag{16b}$$

Then, the Eq. 7a pdf for variable x is modified as follows:

$$p_{x,j} = \frac{1}{(2\pi)^{v/2}\sqrt{\det(V_{x,j})}}\exp\left[-\frac{1}{2}(H_x x - \tilde{x})^T V_{x,j}^{-1}(H_x x - \tilde{x})\right] \qquad j = 1, 2 \tag{17}$$

where $V_{x,j}$ is the diagonal matrix containing the variances $\sigma_{x,j}^2$. A similar expression for $p_{y,j}$ may be written that allows the global log-likelihood function to be expressed:

$$\Phi = \log[wp_{x,1} + (1-w)p_{x,2}][wp_{y,1} + (1-w)p_{y,2}] \tag{18}$$

Following the same step as previously, the Lagrange function associated with the minimization of Eq. 18, subjected to the constraints of Eq. 6, can be explained by Eq. 10. Then, by direct derivative of this Lagrange function, the optimality equations may be expressed as

$$H_x^T W_{\hat{x}}^{-1}(H_x\hat{x} - \tilde{x}) + A^T\lambda + (A \otimes \hat{y})^T\mu = 0 \tag{19a}$$

$$H_y^T W_{\hat{y}}^{-1}(H_y\hat{y} - \tilde{y}) + (A \otimes \hat{x})^T\mu = 0 \tag{19b}$$

$$A\hat{x} = 0 \tag{19c}$$

$$A(\hat{x} \otimes \hat{y}) = 0 \tag{19d}$$

In these last expressions, the weight matrices $W_{\hat{x}}$ and $W_{\hat{y}}$ were already defined in Eqs. 12a and 12b. As in the first section, observability for x and y is needed[4]; in that case, Eqs. 19a–19d can be transformed into the following implicit system:

$$\hat{x} = [G_{\hat{x}} - G_{\hat{x}}A^T(AG_{\hat{x}}A^T)^{-1}AG_{\hat{x}}] \times [H_x^T W_{\hat{x}}^{-1}\tilde{x} - A_{\hat{y}}^T(A_{\hat{x}}G_{\hat{y}}A_{\hat{x}}^T)^{-1}A_{\hat{x}}G_{\hat{y}}H_y^T W_{\hat{x}}^{-1}\tilde{y}] \quad (20a)$$

$$\hat{y} = [G_{\hat{y}} - G_{\hat{y}}A_{\hat{x}}^T(A_{\hat{x}}G_{\hat{y}}A_{\hat{x}}^T)^{-1}A_{\hat{x}}G_{\hat{y}}]H_y^T W_{\hat{y}}^{-1}\tilde{y} \quad (20b)$$

$$G_{\hat{x}} = (H_x^T W_{\hat{x}}^{-1}H_x + A^T A)^{-1} \quad (20c)$$

$$G_{\hat{y}} = (H_y^T W_{\hat{y}}^{-1}H_y + A_{\hat{x}}^T A_{\hat{x}})^{-1} \quad (20d)$$

Comparing the structures of Eq. 20 and Eq. 13 allows us to use the iterative scheme of the third section. Thus, the same estimation scheme for $\hat{x}$ and $\hat{y}$ may be applied when either all or a part of the variables are measured.

Bounded data reconciliation

The previous approach to data reconciliation ensures that the estimates of process variables satisfy the total mass (linear constraint) and partial mass balances (bilinear constraint). However, additional constraints, such as nonnegativity restrictions on the flow rates or known upper and lower bounds on the process variables, are not taken into account.[7] In this section we propose a procedure for incorporating bounds on process variables in the data reconciliation problem itself. These constraints may be either natural, that is, resulting from physical definitions of variables (a flow rate is positive, a volumic concentration is positive and less that a maximum value) or resulting from empirical knowledge (the operator knows that such flow rate must be greater than a given threshold). The proposed procedure is developed in the context of linear constraints only and when all the process variables are measured. Its generalization to more complex cases (bilinear constraints and/or partial measurements) is straightforward. In the following, the variable x is constrained to belong to an interval

$$\underline{x} \le x \le \bar{x} \quad (21)$$

An elegant way to take into account such a constraint consists in using a Bayesian estimator. Let us recall the Bayes rule expressing the posterior probability density function (that is, the conditional density of x given its measurement $\tilde{x}$)

$$p(x|\tilde{x}) = \frac{p(\tilde{x}|x)p(x)}{p(\tilde{x})} \quad (22)$$

In this last expression, $p(\tilde{x}\,|\,x)$ denotes the conditional pdf of the observation given the true value of x, and $p(x)$ is the prior external pdf of x that can be incorporated into the estimation problem. Given that the denominator of Eq. 22 is a constant, the posterior density reduces to $p(x\,|\,\tilde{x}) \propto p(\tilde{x}\,|\,x)p(x)$. To facilitate the estimate computation, continuous pdf values were used both for the likelihood function (as in the previous section) and the prior pdf, which was modeled by

$$p(x) = \frac{1}{2}\left[\tanh\left(\frac{x-\underline{x}}{r}\right) - \tanh\left(\frac{x-\bar{x}}{r}\right)\right] \quad (23)$$

Table 1. Process Equations

Node	Equations	
1	$x_1 - x_2 - x_4 = 0$	$x_1y_1 - x_2y_2 - x_4y_4 = 0$
2	$x_2 - x_3 - x_{11} = 0$	$x_2y_2 - x_3y_3 - x_{11}y_{11} = 0$
3	$x_3 - x_4 - x_5 = 0$	$x_3y_3 - x_4y_4 - x_5y_5 = 0$
4	$x_5 - x_6 - x_{10} = 0$	$x_5y_5 - x_6y_6 - x_{10}y_{10} = 0$
5	$x_6 - x_7 - x_8 = 0$	$x_6y_6 - x_7y_7 - x_8y_8 = 0$
6	$x_7 - x_9 - x_{10} = 0$	$x_7y_7 - x_9y_9 - x_{10}y_{10} = 0$
7	$x_{11} - x_{12} - x_{13} - x_{16} = 0$	$x_{11}y_{11} - x_{12}y_{12} - x_{13}y_{13} - x_{16}y_{16} = 0$
8	$x_{12} - x_{13} - x_{14} = 0$	$x_{12}y_{12} - x_{13}y_{13} - x_{14}y_{14} = 0$
9	$x_{14} - x_{15} - x_{16} = 0$	$x_{14}y_{14} - x_{15}y_{15} - x_{16}y_{16} = 0$

and where the smaller r is chosen, the better the approximation of Eq. 21 is obtained. Thus, in our case, the prior log-pdf of x is expressed as

$$\log p(\tilde{x}|x) = \log\left[\frac{w}{\sqrt{\det(V_{x,1})}}\exp\left(-\frac{1}{2}\|x-\tilde{x}\|^2_{V_{x,1}^{-1}}\right) + \frac{1-w}{\sqrt{\det(V_{x,2})}}\exp\left(-\frac{1}{2}\|x-\tilde{x}\|^2_{V_{x,2}^{-1}}\right)\right] + \log\left[\tanh\left(\frac{x-\underline{x}}{r}\right) - \tanh\left(\frac{x-\bar{x}}{r}\right)\right] - \log 2 - \log[(2\pi)^{v/2}] \quad (24)$$

The associated Lagrange function becomes

$$\mathcal{L} = \log p(\tilde{x}|x) + \lambda^T Ax \quad (25)$$

The derivative of $\mathcal{L}$ with respect to x gives

$$\frac{\partial p(\tilde{x}|x)}{\partial x} = W_x^{-1}[x - (\tilde{x} + W_x h_x)] + A^T\lambda \quad (26)$$

with

$$W_x^{-1} = \frac{wp_{x,1}V_{x,1}^{-1} + (1-w)p_{x,2}V_{x,2}^{-1}}{wp_{x,1} + (1-w)p_{x,2}} \quad (27a)$$

$$h_x = \frac{1}{r}\left[\tanh\left(\frac{x-\underline{x}}{r}\right) + \tanh\left(\frac{x-\bar{x}}{r}\right)\right] \quad (27b)$$

A consideration of Eq. 26 clearly shows that the bounds are taken into account with respect to the quantity h_x. Thus, results of section 3 can be applied when substituting W_x by its new definition (Eq. 27a) and $\tilde{x}$ by $\tilde{x} + W_x h_x$. The two proposed extensions concerning partial measurements and bounded estimations extend the use of the reconciliation procedure. Moreover, it is possible to use them simultaneously.

Example and Discussion

The method described in the previous sections has been successfully applied to the flowsheet given in Smith and Lewis[21] and to simulated processes. Here, we present the results obtained from the system described by equations assembled in Table 1; the given equations model a plant with 9 units and 15 streams, each stream being characterized by a total

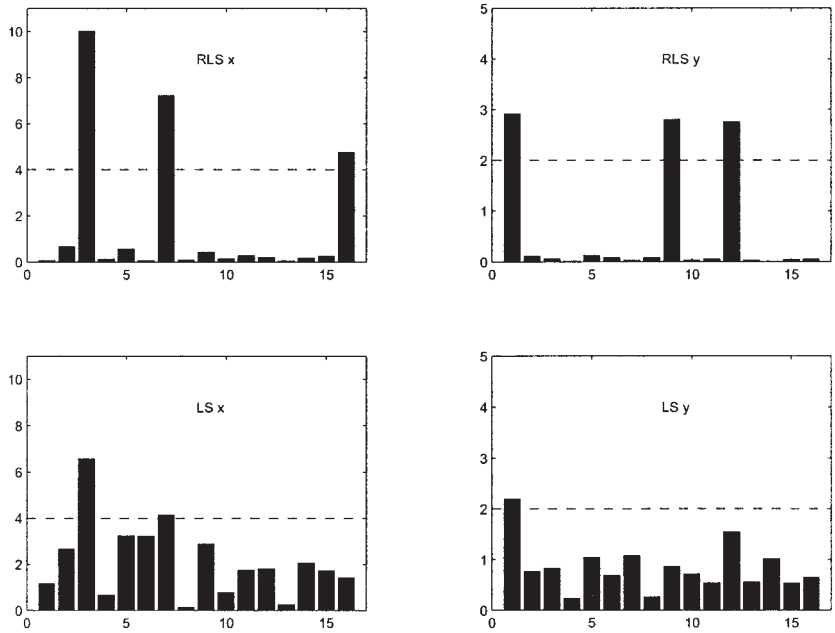

Figure 1. Corrective terms for RLS and LS.

Dashed lines correspond to detection thresholds.

flow rate denoted x and a concentration in a particular chemical or mineral species denoted by y. Random errors were added to the 16 variables, whereas some gross errors corrupt some of them. The simulation results were obtained using the following parameter values: $w = 0.9$, $r = 0.0025$, $\sigma^2_{x,1} = 100x$, $\sigma^2_{x,2} = 0.1x$, $\sigma^2_{y,1} = 10^{-5}y$, and $\sigma_{y,2} = 10^{-8}y$ (the variance of the pdf was chosen proportional to the measurements with a factor of 1000 to express the contamination).

Comparison of the proposed robust least-square (RLS) algorithm with the classical least-square (LS) algorithm is now provided for the studied example. The x data were corrupted with gross errors on components 3, 7, and 16 with respective magnitudes 10, 8, and 5, whereas the components 1, 9, and 12 of the y data were affected by gross errors of same magnitude equal to 3. In the following, the detection threshold for gross errors have been fixed, respectively, to 4 and 2 for x and y data (see the corresponding dashed lines in Figure 1).

In a first test run, observation of variable x_5 and y_2 are missing and bounds on variables are defined (for convenience, only the lower bound $\underline{x}$ on x is given in Table 2). With the chosen values, only the flow rate estimation of the ninth stream was bounded. Without ambiguity, all the gross errors were detected and isolated with RLS, which is not the case with LS. Indeed, in Table 2, the true data are given in columns 3 and 7 and their respective measurements in columns 4 and 8. Columns 5 and 6 for the x variable (columns 9 and 10 for the y variable, respectively) show the estimations obtained with RLS and LS methods. In this table, bold characters indicate the true values, the RLS and the LS estimations for the variables contaminated by gross errors. Analyzing the corrective terms

Table 2. Measurements and Estimations*

Variable	Bound $\underline{x}$	True Data x	Meas. $\tilde{x}$	RLS Est. $\hat{x}$	LS Est. $\hat{x}$	True Data y	Meas. $\tilde{y}$	RLS Est. $\hat{y}$	LS Est. $\hat{y}$
1	50	57.72	57.74	58.35	59.47	**6.23**	8.96	**6.36**	**7.08**
2	50	67.71	67.05	66.61	68.29			7.04	7.67
3	50	**52.98**	63.91	**53.74**	**57.07**	6.37	6.60	6.52	7.26
4	5	7.99	7.94	8.25	8.82	11.65	11.84	11.86	11.56
5	10			45.49	48.25	5.43	5.45	5.55	6.45
6	50	55.71	55.89	55.97	59.36	6.40	6.26	6.43	7.03
7	30	**32.13**	39.49	**32.49**	**35.50**	7.24	7.10	7.19	8.32
8	20	23.58	23.44	23.48	23.86	5.26	5.42	5.37	5.10
9	22	21.40	21.35	22.00	24.39	**5.62**	8.78	**5.74**	**7.78**
10	5	10.73	10.45	10.48	11.11	10.47	10.31	10.23	9.52
11	5	12.73	13.03	12.86	11.22	9.07	9.25	9.22	9.75
12	5	17.05	16.56	16.76	18.59	**8.80**	11.85	**8.69**	**10.01**
13	1	2.42	2.38	2.32	2.65	22.02	22.81	21.79	21.16
14	5	19.47	18.90	19.08	21.24	10.45	10.23	10.29	11.40
15	5	12.73	13.07	12.86	11.22	9.07	9.13	9.22	9.75
16	0	**6.74**	11.63	**6.22**	**10.02**	13.05	12.46	12.48	13.24

*For the y variable, values are multiplied by 100.

Table 3. Percentages of Gross Error Detection

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x, RLS	8.2	8.4	**99.4**	0.0	2.5	6.8	**86.2**	0.2	1.0	0.0	0.0	0.8	0.0	2.0	0.0	**95.8**
y, RLS	**99.6**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**86.6**	0.0	0.0	**97.2**	0.0	0.0	0.0	0.0
x, LS	8.4	10.2	**99.2**	0.0	4.0	8.5	**76.0**	0.0	0.0	0.0	0.0	7.2	0.0	0.0	0.0	**0.0**
y, LS	**97.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.2	**0.0**	0.0	0.0	**13.4**	0.0	0.0	0.0	0.0

for RLS estimator clearly allows us to suspect variables 3, 7, and 16 for being contaminated by a gross error. Such conclusions are more difficult to express with the LS estimator. Moreover, it is instructive to examine the corrective terms affecting the variables, for both those free of gross error and those affected by gross errors. It is clear that the LS approach leads to scattering the corrective terms on all the variables and not only on those corrupted by gross errors. On the contrary, the RLS approach mainly affects the corrective terms on the data that have been subjected to gross errors.

The second test is designed to evaluate the performance of the proposed approach by conducting a Monte Carlo experiment with 500 simulations. Figure 1 more clearly shows the estimation errors for both the LS and the RLS methods (on each graph, horizontal and vertical axes are scaled, respectively, with the number of the data and the magnitude of the absolute estimation error). For each simulation, the same set of gross errors as previously defined is used, but the generation of random noise is renewed for each simulation. Thus, for the whole simulation set, we have generated 8000 data for the x variable (among them 1500 gross errors) and the same number for the y variable (among them 1500 gross errors). Detection of gross errors is performed by comparing the corrective terms $\hat{x} - \tilde{x}$ and $\hat{y} - \tilde{y}$ with a given threshold (4 and 2 for x and y, respectively). Table 3 presents, for each variable, the number of gross errors that have been detected (the results are expressed as a percentage of the total number of runs). The first row indicates the number of the variable (x or y), rows 2 and 3 show the results obtained with our approach, whereas rows 4 and 5 relate the results given by the standard LS approach. The percentages in bold concern the measurements that have been corrupted by gross errors; for example, gross errors on x have been detected on variables 3, 7, and 16 with successful percentages of 99.4, 86.2, and 95.8. These results have to be compared advantageously with 99.2, 76.0, and 0, the last score indicating that gross error on variable 16 has never been detected with the LS approach. For the other variables, the reader should appreciate the level of false detection. An analogous conclusion may be drawn with the y variable.

To complete this analysis, let us observe the magnitudes of the corrective terms: with the RLS approach, the corrections for x variable of streams 3, 7, and 16 are approximately 9.47, 8.25, and 4.40, which can be satisfactorily compared with the three gross error magnitudes 10, 8, and 5. For the y variable, corrections are 2.92, 2.80, and 2.83 characterized by gross error magnitudes of 3, 3, and 3. Of course, detection of gross errors depends on the chosen detection threshold. Here, a fixed cutoff has been selected and the given detection statistics depend on that cutoff. In fact, when analyzing Figure 1, there are a number of thresholds giving the same results. This situation is actually attributed to the improvement of contrast between the magnitudes of the corrections made to the faulty measurements and the others.

Conclusion

To deal with the issues of gross error influence on data estimation, this paper has presented a robust approach for data reconciliation using a cost function that is less sensitive to the outlying observations than that of least squares. Although we consider only the class of static linear and bilinear systems, the proposed approach covers many applications in the field of chemical and mineralogical engineering. As a perspective of development of robust reconciliation strategies, there is a need for taking into account the model uncertainties and optimizing the balancing parameter w that define the compromise between noise and gross error distribution. Moreover, there is some potential for adapting the strategy to dynamical linear systems.

Literature Cited

1. Narasimhan S, Jordache C. *Data Reconciliation and Gross Error Detection*. Houston, TX: Gulf Publishing; 2000.
2. Ragot J, Darouach M, Maquin D, Bloch G. *Validation de données et diagnostic*. Paris, France: Hermès; 1990.
3. Romagnoli J, Sanchez M. *Data Processing and Reconciliation for Chemical Process*. New York, NY: Academic Press; 2000.
4. Bagajewicz MJ. *Process Plant Instrumentation: Design and Upgrade*. Lancaster, PA: Technomic; 2000.
5. Albuquerque J, Biegler LT. Data reconciliation and gross errors detection for dynamic systems. *AIChE J*. 1996;42:2841-2856.
6. Jordache C, Ternet D, Brown S. Efficient gross error elimination methods for rigorous on-line optimization. Proc of Escape 11, Kolding, Denmark; 2001.
7. Narasimhan S, Harikumar P. A method to incorporate bounds in data reconciliation and gross error detection. *Comput Chem Eng*. 1993;17:115-1120.
8. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stohel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York, NJ: Wiley; 1986.
9. Goodall C. M-estimators of location: An outline of the theory. In: Hoaglin D, Mosteller F, Tukey JW, eds. *Understanding Robust and Explanatory Data Analysis*. New York, NY: Wiley; 1983:339-403.
10. Hubert P. *Robust Statistics*. New York, NY: Wiley; 1981.
11. Tjoa IB, Biegler LT. Simultaneous strategy for data reconciliation and gross error analysis. *Comput Chem Eng*. 1991;15:679-689.
12. Singhal A, Seborg DE. Dynamic data rectification using the expectation maximisation algorithm. Technical report PC-101. Berkeley, CA: Dept. of Chemical Engineering, University of California; 2000.
13. Wang D, Romagnoli J. Robust data reconciliation based on a generalized objective function. Proc of the 15th IFAC Triennal World Congress, Barcelona, Spain; 2002.
14. Soderstrom T, Young R, Russo L, Edgar TF. Industrial application of a large-scale dynamic data reconciliation strategy. *Ind Eng Chem Res*. 2000;39:1683-1693.
15. Ghosh-Dastider B, Schafer JL. Outlier detection and editing procedure for continuous multivariate data. Working paper 07. Princeton, NJ: Office of Population Research, Princeton University; 2003.
16. Mah RSH, Stanley GM, Downing D. Reconciliation and rectification

of process flow and inventory data. *Ind Eng Chem Proc Des Dev*. 1976;15:175-183.
17. Hodouin D, Flament F. New developments in material balance calculations for mineral processing industry. Proc of the Society of Mining Engineers Annual Meeting, Las Vegas, NV, February 27–March 2; 1989.
18. Crowe CM. Data reconciliation—Progress and challenges. *J Process Control*. 1996;6:89-98.
19. Maquin D, Bloch G, Ragot J. Data reconciliation for measurements. *Eur J Diagn Safety Automat*. 1991;1:145-181.
20. Tukey JW. A survey of sampling from contaminated distributions. In: Olkin I, Ghury SG, Hoeffding W, Madow WG, Mann HB, eds. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford, CA: Stanford Univ. Press; 1960;448-485.
21. Smith HW, Lewis CL. Computer adjustment of metallurgical balances. *Can Inst Mining Metall Bull*. 1973;66:97-100.

Manuscript received Mar. 6, 2004, and revision received Sep. 16, 2004.