Data reconciliation : a robust approach using contaminated distribution. Application in mineral processing for multicomponent products

Moustapha Alhaj-Dibo, Didier Maquin, José Ragot
Centre de Recherche en Automatique de Nancy, CNRS UMR 7039
Institut National Polytechnique de Lorraine
2, Avenue de la forêt de Haye, 54516 Vandœuvre-les-Nancy Cedex, FRANCE
{Moustapha.Alhajdibo, Didier.Maquin, José.Ragot}@ensem.inpl-nancy.fr

Key words: mass balance, data reconciliation, robust estimation, gross errors, outliers.

Abstract

On-line optimisation provides a means for maintaining a process around its optimum operating plant. An important component of optimisation relies in data reconciliation which is used for obtaining consistent data. On a mathematical point of view, the formulation is generally based on the assumption that the measurement errors have normally pdf with zero mean. Unfortunately, in the presence of gross errors, all of the adjustments are greatly affected by such biases and would not be considered as reliable indicators of the state of the process. This paper proposes a data reconciliation strategy that deals with the presence of such gross errors. Application to size flowrates and concentration data in mineral processing is provided.

x, y	process variables
$\tilde{x}, \tilde{y}$	process measurements
$\hat{x}, \hat{y}$	estimates
σ	standard deviation
p	probability density function
A	incidence matrix
v	number of streams
n	number of nodes
q	number of concentrations
W_x, W_y	weighting matrix
w	balancing factor
V	variance matrix
λ, μ	Lagrange parameters

Table 1: List of symbols

1 Introduction

The problem of obtaining reliable estimates of the state of a process is a fundamental objective, these estimates being used to understand the process behaviour. For that purpose, a wide variety of techniques has been developed to perform what is currently known as data reconciliation [Mah, 76], [Maquin, 91]. Data reconciliation, which is sometimes referred too as mass and energy balance equilibration, is the adjustment of a set of data so the quantities derived from the data obey physical laws such as material and energy conservation. Since the pionner works devoted to the so-called data rectification, the scope of research has expanded to cover other fields such as data redundancy analysis, system observability, optimal sensor positionning, sensor reliability, errors characterization, measurement variance estimation. Many applications are related in scientific papers involving various fields in process engineering [Yi, 02], [Singh, 01], [Heyen, 99].

Unfortunately, the measurement collected on the process may be unknowingly corrupted by gross errors. As a result, the data reconciliation procedure can give rise to absurd results and, in particular, the estimated variables will be corrupted by this bias. Several schemes have been suggested to cope with the corruption of normal assumption of the errors, for static system [Narasimhan, 89], [Arora, 01] and also for dynamic systems [Abu-el-zeet, 01]. Methods to include bounds in process variables to improve gross errors detection have been developed. One major disadvantage of these methods is that they give rise to situations that it may impossible to estimate all the variable by using only a subset of the remaining free gross errors measurements. Alternative approach using constraints both on the estimates and the balance residual equations has been developped for linear system [Ragot, 99], [Maquin, 04]. There is also an important class of robust estimators whose influence function are bounded and finit allowing to reject outliers

[Huber, 81], [Hampel, 86]. Another approach is to take into account the non ideality of the measurement error distribution by using an objective function constructed on contaminated error distribution. In the following, we adopt and develop this idea for the data reconciliation problem.

Section 2 will be devoted to recall the background of data reconciliation. In section 3, robust data reconciliation is developped and will be illustrated through an academic example in section 4.

2 Data reconciliation background

The classical general data reconciliation problem [Mah, 76], [Hodouin, 89] [Crowe, 96], deals with a weighted least squares minimisation of the measurement adjustments subject to the model constraints. Indeed the model process equations are taken as linear for sake of simplicity :

$$Ax = 0, \quad A \in \mathbb{R}^{n.v}, \quad x \in \mathbb{R}^{v} \tag{1}$$

where x, with components x_i is the state of the process. The measurement devices give the information :

$$\tilde{x} = x + \epsilon, \quad p(\epsilon) \propto N(0, V) \tag{2}$$

where $\epsilon \in R^n$ is a vector of random errors characterised by variance matrix V and p is the normal probability distribution (pdf). In the least square sense, the well-known solution of this problem is $\hat{x} = (I - VA^T(AVA^T)^{-1}A)x$ [Maquin, 1991]. In fact, the method doesn't work in any situation, the main drawback being the contamination of all estimated values by the outliers. For that reason robust estimators could be preferred, robustness being the ability to ignore the contribution of extreme data i.e. such as gross errors. There are two approaches to deal with outliers. The first one consist to sequentially detect, localise and suppress the data which are contaminated and after to reconcile the remaining data. The second approach is global and reconcile the data without a preliminary classification; in fact, weights in the reconciliation procedure are automatically adjust in order to minimise the influence of the abnormal data. In the rest of the paper, we only focuse on this last strategy.

3 Robust data validation. The linear case.

If the measurements contain random outliers, then a single pdf described as in (2) cannot account for the high variance of the outliers. To overcome this problem let us assume that measurement noise is sampled from two pdf, one having a small variance representing regular noise and the other having a large variance representing outliers [Wang, 02], [Ghosh, 03]. Thus, for each observation $\tilde{x}_i$, we define the two following pdf and the so-called contaminated pdf:

$$p_{j,i}(x_i \mid \tilde{x}_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_j} exp\left(-\frac{1}{2}\left(\frac{x_i - \tilde{x}_i}{\sigma_j}\right)^2\right) \tag{3}$$

$$p(x_i \mid \tilde{x}_i, \theta) = wp_{1,i} + (1-w)p_{2,i} \quad 0 \le w \le 1 \tag{4}$$

allowing to define the log-likelihood function of the measurement set:

$$\Phi = log \prod_{i=1}^{v} p(x_i \mid \tilde{x}_i, \theta) \tag{5}$$

Minimising (5) in respect to x gives the estimate $\hat{x}$:

$$\hat{x} = (I - W_x A^T (AW_x A^T)^{-1} A)\tilde{x} \tag{6a}$$

$$W_x^{-1} = \underset{i\,=\,1..v}{diag} \left(\frac{\frac{w}{\sigma_1^2}\hat{p}_{1,i} + \frac{1-w}{\sigma_2^2}\hat{p}_{2,i}}{w\hat{p}_{1,i} + (1-w)\hat{p}_{2,i}}\right) \tag{6b}$$

$$\hat{p}_{j,i} = \frac{1}{\sqrt{2\pi}\sigma_j} exp\left(-\frac{1}{2}\left(\frac{\hat{x}_i - \tilde{x}_i}{\sigma_j}\right)^2\right) \tag{6c}$$

where the diag operator allow to define a diagonal matrix from the elements (pointed by i) of a vector. Thus system (6) is clearly non linear and we suggest the following direct iterative scheme:

$$x^{(0)} = \tilde{x} \tag{7a}$$

$$p_{j,i}^{(k)} = \frac{1}{\sqrt{2\pi}\sigma_j} exp\left(-\frac{1}{2}\left(\frac{x_j^{(k)} - \tilde{x}_j}{\sigma_j}\right)^2\right) \tag{7b}$$

$$(W_x^{(k)})^{-1} = \underset{i\,=\,1..v}{diag} \left(\frac{\frac{w}{\sigma_1^2}\hat{p}_{1,i}^{(k)} + \frac{1-w}{\sigma_2^2}\hat{p}_{2,i}^{(k)}}{w\hat{p}_{1,i}^{(k)} + (1-w)\hat{p}_{2,i}^{(k)}}\right) \tag{7c}$$

$$\hat{x}^{(k+1)} = \left(I - W_x^{(k)} A^T (AW_x^{(k)} A^T)^{-1} A\right)\tilde{x} \tag{7d}$$

A stopping criterion must be chosen for implementing the algorithm. For sake of simplicity, the proof for the local convergence of the algorithm is omitted.

In order to appreciate how the weight W, which should be compared to an influence function as explained in [Hampel, 86], are able to reject the data contaminated by gross errors, figure 1 show the graph of the function:

$$g(u) = \frac{\frac{w}{\sigma_1^2}p_1 + \frac{1-w}{\sigma_2^2}p_2}{wp_1 + (1-w)p_2}$$

$$p_1 = \frac{1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{1}{2}(\frac{u}{\sigma_1})^2\right)$$

$$p_2 = \frac{1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{1}{2}(\frac{u}{\sigma_2})^2\right)$$

where $\sigma_1 = 0.5$ and $\sigma_2 = 2$ and where w take the indicated values. For a better comparison, the graphs have been normalized, i.e. we have represented $\overline{g}(u) = g(u)/g(0)$. For $w = 1$ we naturally obtain a constant weight; thus all the data are equallly weighted and, in particular, the optimisation criterion will be sensitive to large magnitude of data, i.e. to outliers. Taking $w = 0.02$ reduces the influence of outliers since the weight decreases from 1 for data around the origine to 0.63 for data with large magnitude. Indeed, with the non restrictive hypothesis $\sigma_2 > \sigma_1$, for large values of u, the weighting function $\overline{g}(u)$ can be approximated by the non zero value:

$$g_a(u) = \frac{1}{\sigma_1^2} \frac{1 + (\frac{1-w}{w})(\frac{\sigma_1}{\sigma_2})^3}{1 + (\frac{1-w}{w})(\frac{\sigma_1}{\sigma_2})}$$

where for small values of u the approximation is $g_b(u) = 1$. Thus, it is possible to adjust σ_1 and σ_2 such that the large values of u would have a small influence on the criterion Φ.

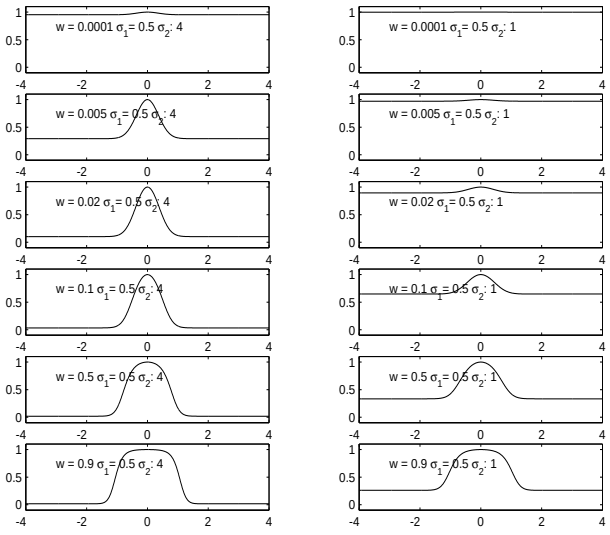

Figure 1: Influence function

4 Extension to bilinear systems

We consider now the case of a process characterised by two types of variables : macroscopic variables such as flowrates x and microscopic variables such concentrations y. Moreover, we will consider several species and therefore several concentrations noted $y_c, c = 1..q$. If the measurements contain random outliers, then a single pdf described as in (2) cannot account for the high variance of the outliers. To overcome this problem let us assume that measurement noise is sampled from two pdf, one having a small variance representing regular noise and the other having a large variance representing outliers. In order to simplify the presentation, each measurement x_i (resp. y_i) are assumed to have the same normal $\sigma_{x,1}$ (resp. $\sigma_{y,1}$) and abnormal $\sigma_{x,2}$ (resp. $\sigma_{y,2}$) standard-deviation. This hypothesis will be withdrawn later. Thus, for each observation $\tilde{x}_i$ and $\tilde{y}_{c,i}$, we define the following pdf:

$$p(x_i|\tilde{x}_i, \sigma_{x,j}) = \frac{1}{\sqrt{2\pi}\sigma_{x,j}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \tilde{x}_i}{\sigma_{x,j}}\right)^2\right) \quad (8a)$$

$$p(y_{c,i}|\tilde{y}_{c,i}, \sigma_{y_{c,j}}) = \frac{1}{\sqrt{2\pi}\sigma_{y_{c,j}}} \exp\left(-\frac{1}{2}\left(\frac{y_{c,i} - \tilde{y}_{c,i}}{\sigma_{y_{c,j}}}\right)^2\right) \quad (8b)$$

with $j = 1, 2, i = 1..v, c = 1..q$. In the rest of the paper, we adopt the shortening notation $p_{x,j,i}$ and $p_{y_c,j,i}$ respectively for $p(x_i|\tilde{x}_i, \sigma_{x,j})$, and $p(y_{c,i}|\tilde{y}_{c,i}, \sigma_{y_c,j})$ where indexes i and j are respectively used to point the number of data and the number of the distribution. Then, the combination of these two pdf (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 occurence of outliers:

$$p_{x,i} = wp_{x,1,i} + (1-w)p_{x,2,i} \quad i = 1..v \quad (9a)$$

$$p_{y_c,i} = wp_{y_c,1,i} + (1-w)p_{y_c,2,i} \quad i = 1..v \quad (9b)$$

Assuming independance of the measurements allows to define the global log-likelihood function:

$$\Phi = log \prod_{i=1}^{v} p_{x,i} \prod_{c=1}^{q} p_{y_c,i} \quad (10)$$

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

$$L = \Phi + \lambda^T Ax + \sum_{c=1}^{q} \mu_c^T A(x \otimes y_c) \quad (11)$$

in which the parameters λ and μ_c allow to take into account the mass balance constraints for total flowrate and partial flowrate (for that last one the operator $\otimes$ is used to perform the element by element product of two vectors). The stationarity conditions of (11) are expressed (the estimations are now noted $\hat{x}$ and $\hat{y}_c$):

$$W_{\hat{x}}^{-1}(\hat{x} - \tilde{x}) + A^T\lambda + \sum_{c=1}^{q}(A \otimes \hat{y}_c)^T\mu_c = 0 \quad (12a)$$

$$W_{\hat{y}_c}{}^{-1}(\hat{y}_c - \tilde{y}_c) + (A \otimes \hat{x})^T\mu = 0 \quad (12b)$$

$$A(\hat{x} \otimes \hat{y}_c) = 0 \quad (12c)$$

where the weighting matrices $W_{\hat{x}}$ and $W_{\hat{y}_c}$ are defined by:

$$W_{\hat{x}}^{-1} = \underset{i=1..v}{diag} \left(\frac{\frac{wp_{x,1,i}}{\sigma_{x,1}^2} + \frac{(1-w)p_{x,2,i}}{\sigma_{x,2}^2}}{wp_{x,1,i} + (1-w)p_{x,2,i}}\right) \quad (13a)$$

$$W_{\hat{y}_c}^{-1} = \underset{i=1..v}{diag} \left(\frac{\frac{wp_{y_c,1,i}}{\sigma_{y_c,1}^2} + \frac{(1-w)p_{y_c,2,i}}{\sigma_{y_c,2}^2}}{wp_{y_c,1,i} + (1-w)p_{y_c,2,i}}\right) \quad (13b)$$

Notice that if each measurement x_i (resp. y_i) has a particular standard-deviation, formulas (13a) and (13b) still hold by replacing the parameters $\sigma_{x,1}$ and $\sigma_{x,2}$ (resp. $\sigma_{y_c,1}$ and $\sigma_{y_c,2}$) by $\sigma_{x,1,i}$ and $\sigma_{x,2,i}$ (resp. $\sigma_{y_c,1,i}$ and $\sigma_{y_c,2,i}$). Using shortening notations $A_x = A\,diag(x)$ and $A_y = A\,diag(y)$, system (12) may be directly solved and the solution may be expressed:

$$\hat{x} = (I - W_{\hat{x}} A^T (A W_{\hat{x}} A^T)^{-1} A)...$$

$$...(\tilde{x} - W_{\hat{x}} \sum_{c=1}^{q} A_{\hat{y}_c}^T (A_{\hat{x}} W_{\hat{y}_c} A_{\hat{x}}^T)^{-1} A_{\hat{x}} \tilde{y}_c) \qquad (14a)$$

$$\hat{y}_c = (I - W_{\hat{y}_c} A_{\hat{x}}^T (A_{\hat{x}} W_{\hat{y}_c} A_{\hat{x}}^T)^{-1} A_{\hat{x}}) \tilde{y}_c \qquad (14b)$$

System (14) is clearly non linear with regard to the unknown $\hat{x}$ and $\hat{y}_c$, the weight $W_{\hat{x}}$ and $W_{\hat{y}_c}$ depending on the pdf (8) which themselves depend on the $\hat{x}$ and $\hat{y}_c$ estimations (**??**) and (14b). In fact (14) is an implicit system in respect to the estimates $\hat{x}$ and $\hat{y}_c$ for which we suggest the following iterative scheme:

Step 1: initialisation
$k = 0$
$\hat{x}^{(k)} = \tilde{x} \quad \hat{y}_c^{(k)} = \tilde{y}_c$
choose w (for example between 0.9 and 0.99)
adjust $\sigma_{x,1}$ and $\sigma_{y_c,1}$ from an a priori knowledge about the noise distribution
adjust $\sigma_{x,2}$ and $\sigma_{y_c,2}$ from an a priori knowledge about the gross error distribution.
Step 2: estimation
Compute the quantities (for $j = 1, 2$, $i = 1..v$ and $c = 1..q$)

$$p_{\hat{x},j,i}^{(k)} = \frac{1}{\sqrt{2\pi}\sigma_{x,j}} \exp\left(-\frac{1}{2}\left(\frac{\hat{x}_i^{(k)} - \tilde{x}_i}{\sigma_{x,j}}\right)^2\right)$$

$$p_{\hat{y}_c,j,i}^{(k)} = \frac{1}{\sqrt{2\pi}\sigma_{y_c,j}} \exp\left(-\frac{1}{2}\left(\frac{\hat{y}_{c,i}^{(k)} - \tilde{y_c}_i}{\sigma_{y_c,j}}\right)^2\right)$$

$$W_{\hat{x}}^{-1} = \underset{i=1..v}{diag} \left(\frac{\frac{w p_{\hat{x},1,i}^{(k)}}{\sigma_{x,1}^2} + \frac{(1-w) p_{\hat{x},2,i}^{(k)}}{\sigma_{x,2}^2}}{w p_{\hat{x},1,i}^{(k)} + (1-w) p_{\hat{x},2,i}^{(k)}}\right)$$

$$W_{\hat{y}_c}^{-1} = \underset{i=1..v}{diag} \left(\frac{\frac{w p_{\hat{y}_c,1,i}^{(k)}}{\sigma_{y_c,1}^2} + \frac{(1-w) p_{\hat{y}_c,2,i}^{(k)}}{\sigma_{y_c,2}^2}}{w p_{\hat{y}_c,1,i}^{(k)} + (1-w) p_{\hat{y}_c,2,i}^{(k)}}\right)$$

$$A_{\hat{x}}^{(k)} = A\,diag(\hat{x}^{(k)}) \quad A_{\hat{y}_c}^{(k)} = A\,diag(\hat{y}_c^{(k)})$$

Update the estimation of x and y_c

$$\hat{x}^{(k+1)} = \left(I - W_{\hat{x}}^{(k)} A^T (A W_{\hat{x}}^{(k)} A^T)^{-1} A\right)...$$

$$... \left(\tilde{x} - W_{\hat{x}}^{(k)} \sum_{c=1}^{q} A_{\hat{y}_c}^{(k)T} (A_{\hat{x}}^{(k)} W_{\hat{y}_c}^{(k)} A_{\hat{x}}^{(k)T})^{-1} A_{\hat{x}}^{(k)} \tilde{y}_c\right)$$

$$\hat{y}_c^{(k+1)} = (I - W_{\hat{y}_c}^{(k)} A_{\hat{x}}^{(k)T} (A_{\hat{x}}^{(k)} W_{\hat{y}_c}^{(k)} A_{\hat{x}}^{(k)T})^{-1} A_{\hat{x}}^{(k)}) \tilde{y}_c$$

Step 3: convergence test
Compute an appropriate norm of the corrective terms: $\tau_x^{(k+1)} = \|\hat{x}^{(k+1)} - \tilde{x}\|$ and $\tau_{y_c}^{(k+1)} = \|\hat{y}_c^{(k+1)} - \tilde{y}_c\|$. If the variations $\tau_x^{(k+1)} - \tau_x^{(k+1)}$ and $\tau_{y_c}^{(k+1)} - \tau_{y_c}^{(k+1)}$ are less than a given threshold then stop, else $k = k + 1$ and go to step 2.

Remark : for non linear systems, the initialisation remains a difficult task, convergence of the algorithm being generally sensitive to that choice. In our situation, measurements are a natural choice for initializing the estimates (step 1 of the algorithm). The solution given by classical least squares approach would also provide an acceptable initialization although its sensitivity to gross errors may be sometimes important; the reader should verify that this solution may be obtained by redefining the distributions (9) with $w = 1$.

5 Example and discussion

The method described in section 4 is applied to system depicted by fig.2, for which 16 streams are considered; each stream is characterized by a flowrate and two concentrations. Random errors were added to the 16 variables but the gross errors were added only on some of them.

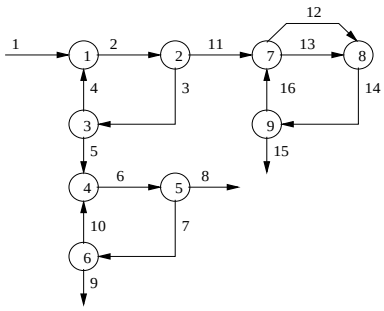

Figure 2: Flowsheet

The performance results are given when three gross errors (with magnitudes of 6, 8 and 8) affect the measurement 3, 7 and 16; simultaneously, gross errors of magnitude 1.5 affect the measurement of the first concentration for streams 1, 9 and 12, and gross errors of magnitudes 4 and 2.5 affect the measurement of the second concentration for streams 4 and 8. Comparison of the proposed robust least square algorithm (RLS) with the classical least squares (LS) algorithm is now provided in table 2.

Columns 2 to 4 relate the row measures, columns 5 to 7 show the estimations obtained with RLS and columns 8 to 10 the estimations obtained with LS ; analysing the estimation errors, for RLS estimator clearly allows to suspect variables 3, 7 and 16 for being contaminated by a gross error. Such conclusion is more difficult to express

	Measurement			RLS estimate			LS estimate		
	x	y_1	y_2	$\hat{x}$	$\hat{y}_1$	$\hat{y}_2$	$\hat{x}$	$\hat{y}_1$	$\hat{y}_2$
1	55.88	**3.93**	3.53	56.50	**2.38**	3.46	57.30	**2.63**	3.58
2	65.31	2.73	3.70	65.07	2.71	3.70	65.80	2.89	4.16
3	**61.68**	2.48	3.54	**52.99**	2.52	3.58	**54.63**	2.73	4.16
4	8.38	4.83	**9.01**	8.57	4.83	**5.30**	8.50	4.62	**8.03**
5	44.15	2.13	3.16	44.42	2.07	3.24	46.13	2.38	3.45
6	55.90	2.45	3.57	55.30	2.52	3.51	56.85	2.67	3.68
7	**39.05**	2.87	3.74	**31.44**	2.87	3.76	**32.97**	3.23	3.49
8	23.90	2.05	**5.72**	23.86	2.05	**3.18**	23.88	1.91	**3.95**
9	20.58	**3.75**	3.34	20.55	**2.09**	3.32	22.25	**2.88**	2.91
10	10.33	4.35	4.60	10.89	4.34	4.60	10.72	3.95	4.69
11	12.40	3.57	4.16	12.08	3.54	4.25	11.17	3.68	4.15
12	17.66	**5.11**	4.32	17.49	**3.47**	4.27	19.85	**3.97**	4.30
13	2.66	8.92	6.91	2.27	8.92	6.90	2.71	8.45	6.90
14	19.38	4.18	4.66	19.76	4.10	4.57	22.56	4.51	4.61
15	12.42	3.45	4.24	12.08	3.54	4.25	11.17	3.68	4.15
16	**14.77**	4.95	5.01	**7.68**	4.98	5.07	**11.39**	5.33	5.06

Table 2: Measurements and estimations

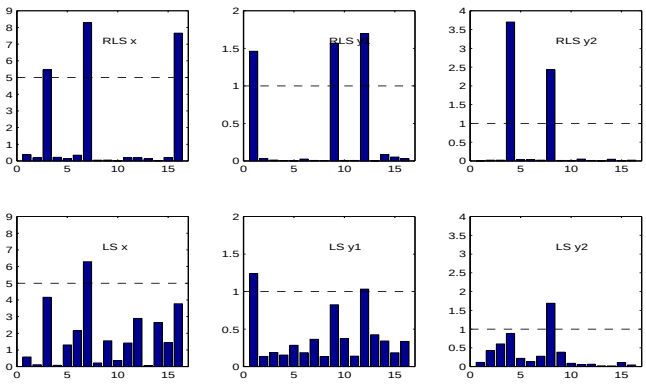

Figure 3: Corrective terms

with LS estimator. Table 3 gives explicitely the corrective terms $\hat{x} - \tilde{x}$ and $\hat{y} - \tilde{y}$ for RLS (row 3) and LS (row 4) approach; for a better comparison, row 2 indicates the true value of the gross error and thus we can appreciate the vicinity of the corrective terms obtained with RLS with the true gross errors.

	x_3	x_7	x_{16}	$y_{1,1}$	$y_{1,9}$	$y_{1,12}$	$y_{2,4}$	$y_{2,8}$
T	6.0	8.0	8.0	1.5	1.5	1.5	4.0	2.5
RLS	8.69	7.61	7.09	1.55	1.69	1.64	3.71	2.54
LS	7.05	6.08	3.38	1.3	0.87	1.14	0.98	1.77

Table 3: Corrective terms

For another data set, figure 3 visualizes more clearly the estimation errors ($\hat{x} - \tilde{x}$ and $\hat{y}_c - y_c$) both for RLS (upper part) and LS (lower part). On each graph, horizontal and vertical axis are respectively scaled with the number of the data and the magnitude of the absolute estimation error; the dashed horizontal line is the threshold chosen to detect abnormal corrective terms. Analysing figure 3 shows two advantages on RLS upon LS approach: first, the corrective terms are more precisely estimated, second, the scattering of the gross errors is less (the corrective terms mainly affect the variables affected by the gross errors and not the others).

Performances of the proposed approach be also analysed when using a great number of data. For that purpose, the same process has been used with different additive random noise on the data, the gross errors being superposed to the same data as previous. 10000 runs have been performed, allowing to enumerate the cases where the gross errors have been correctly detected or not, both for RLS and LS method. Results, expressed in percentage, are shown in table 4. Roughly speaking, for the given example, the ability of gross error detection for RLS is twice of those of LS. This has been confirmed by many other runs involving various distributions of the measurement errors.

The sensitive known parameters w, $\sigma_{x,i}$ and $\sigma_{y_c,j}$ of the

	RLS gross error detection			LS gross error detection		
$Var.$	x	y_1	y_2	x	y_1	y_2
w=0.10	92.5	99.9	91.4	41.4	57.2	55.2

Table 4: Performance of the approach

contaminated distribution may affect determination about outliers and therefore requires special attention. Typically, there is a range of sensible values for these parameters that we can start with. In fact, due to the structure of the function defining the weight, we can reduce these parameters to w, $\sigma_{x,1}/\sigma_{x_2}$ and $\sigma_{y_c,1}/\sigma_{y_c,2}$. Table 5 presents some results of sensitivity, expressed in percentage of correct detection, using the same process. For each result of detection, concerning a particular value of w, $f = \sigma_{x,1}/\sigma_{x_2} = \sigma_{y_c,1}/\sigma_{y_c,2}$, 10000 runs have been performed, each run having the same outliers but specific random noise affecting the measurements. It should be noted that all gross errors may be correctly detected with a proper choice of the parameters w and f, excepted the error on the flowrate x_3. Thus, considering results in table 5, it is relatively easy to adjust manually the parameters w and σ of the method and a "large" range of acceptable values may be found. However, it is also possible to use an adaptive algorithm for the adjusting of these parameters.

w	f	x_3	x_7	x_{16}	$y_{1,1}$	$y_{1,9}$	$y_{1,12}$	$y_{2,4}$	$y_{2,8}$
0.02	25	1	1	1	100	1	1	1	99
0.05	25	1	1	0	100	0	1	1	100
0.30	25	5	45	3	100	0	2	0	100
0.02	50	72	99	99	100	75	100	41	100
0.05	50	77	100	100	100	96	100	59	100
0.30	50	76	100	100	100	100	100	90	100
0.02	75	76	100	100	100	100	100	99	100
0.05	75	76	100	100	100	100	100	100	100
0.30	75	76	100	99	100	100	100	100	100

Table 5: Performance of the approach

6 Conclusion

To deal with the issues of gross errors influence on data estimation, the paper has presented a robust reconciliation approach. For that purpose, we use a cost function which is less sensitive to the outlying observations than that of least squares. The algorithme can handle multiple biaises or outliers at a time and for the given example, 8 outliers have been correctly detected on 48 variables.
The results of reconciliation will clearly depend not only on the data, but also on the model of the process itself. As a perspective of development of robust reconciliation strategies, there is a need for taking account of model uncertainties and optimise the balancing parameter w. Moreover, for process with unknown parameter, it should be important to jointly estimate the reconciled data and the process paramaters.

References

[1] Z.H. Abu-el-zeet, P.D. Roberts, V.M. Becerra. Bias detection and identification in dynamic data reconciliation. Proceeding of the European Control Conference, 2001.

[2] N. Arora. Redescending estimator for data reconciliation and parameter estimation. Computers and Chemical Engineering, 25, p. 1585, 2001.

[3] C. Bazin, D. Hodouin. Importance of covariance in mass balancing of particle size distribution data. Minerals Engineering, 14 (8), 2001.

[4] B. Ghosh-Dastider, J.L. Schafer. Outlier detection and editing procedures for continuous multivariate data. Working paper 2003-07, RAND, Santa Monica OPR, Princeton University, 2003.

[5] C.M. Crowe. Data reconciliation - progress and challenges. Journal of Process Control, 6 (2,3), p. 89-98, 1996.

[6] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stohel. Robust statistic : the approach on influence functions. Wiley, New-York, 1986.

[7] G. Heyen. Industrial Applications of Data Reconciliation: Operating closer to the Limits with improved Design of the Measurement System Workshop on Modelling for Operator Support, Lappeenranta University of Technology, June 22, 1999.

[8] D. Hodouin, F. Flament. New developments in material balance calculations for mineral processing industry. Society of Mining Engineers annual meeting, Las Vegas, February 27 - March 2, 1989.

[9] P.J. Hubert. Robust statistic. John Wiley Sons, New York, 1981.

[10] I. Kim, M.S. Kang, S. Park., T.F. Edgar. Robust data reconciliation and gross error detection: The modified MIMT using NLP. Computers Chem. Engng, 21, 775-782, 1997.

[11] R.S.H. Mah, G.M. Stanley and D. Downing. Reconciliation and rectification of process flow and inventory data, Ind. Eng. Chem., Process Des. Dev., 15 (1), 1976.

[12] D. Maquin, G. Bloch, J. Ragot. Data reconciliation for measurements. Europeen Journal of Diagnosis and Safety in Automation, 1 (2), p. 145-181, 1991.

[13] D. Maquin, J. Ragot. Validation de données issues de systèmes de mesure incertains. To appear in Journal Europen des Systèmes Automatisés, 2004.

[14] S. Narasimhan, R.S.H. Mah. Treatment of general steady state process models in gross error identification. Computers and Chemical Engineering, 13 (7), p. 851-853, 1989.

[15] J. Ragot, D. Maquin, O. Adrot. LMI approach for data reconciliation. 38th Conference of Metallurgists, Symposium Optimization and Control in Minerals, Metals and Materials Processing, Quebec, Canada, August 22-26, 1999.

[16] S.R. Singh, N.K. Mittal, P.K. Sen. A novel data reconciliation and gross error detection tool for the mineral processing industry. Minerals Engineering, 14 (7), 2001.

[17] D. Wang, J.A. Romagnoli. Robust data reconciliation based on a generalized objective function. 15th triennial world congress, Barcelona, 2002.

[18] H.-S. Yi, J. H. Kim and C. Han. Industrial application of gross error estimation and data reconciliation to byproduction gases in iron and steel making plants. International Conference on Control, Automation and Systems, October 16-19, 2002, Muju, Korea.