A Novel Robust Nonlinear Dynamic Data Reconciliation*

GAO Qian(高倩)**, **YAN Weiwu**(阎威武) **and SHAO Huihe**(邵惠鹤)
Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract Outlier in one variable will smear the estimation of other measurements in data reconciliation (DR). In this article, a novel robust method is proposed for nonlinear dynamic data reconciliation, to reduce the influence of outliers on the result of DR. This method introduces a penalty function matrix in a conventional least-square objective function, to assign small weights for outliers and large weights for normal measurements. To avoid the loss of data information, element-wise Mahalanobis distance is proposed, as an improvement on vector-wise distance, to construct a penalty function matrix. The correlation of measurement error is also considered in this article. The method introduces the robust statistical theory into conventional least square estimator by constructing the penalty weight matrix and gets not only good robustness but also simple calculation. Simulation of a continuous stirred tank reactor, verifies the effectiveness of the proposed algorithm.

Keywords nonlinear dynamic data reconciliation, robust, M-estimator, outlier, optimization

1 INTRODUCTION

Data reconciliation (DR) is a task to reconcile process data and to obtain the accurate estimation of measured variables by having them satisfy material and energy balance constraints[1]. The steady-state DR (SSDR) is well-documented and applied satisfactorily to large-scale industrial processes[2—5]. However, chemical processes are intrinsically dynamic and dynamic data reconciliation (DDR) will obtain better state estimations than SSDR[6—9].

Data reconciliation processes usually use a weighted least squares objective function, and are based on the assumption that random errors are distributed normally with zero mean and a known variance. However, weighted least squares objective function can lead to incorrect estimation and severely bias reconciliation when the measured data contains gross errors[10]. To make the estimation insensitive in the presence of a persistence of gross errors, robust approaches have been proposed[11,12]. In these methods the weighed squared residual of the DR objective function is replaced by a robust function. The robust function is usually selected as a convex function, to ensure that the solution is unique. The influence function, the derivative of robust function with respect to the process variable measurements, gives a weight approach to zero, to a high value residual, and compensates for the effects that have the residuals on the estimator[13,14]. Comparison of least square estimator and robust M-estimator shows the latter is superior in detecting outliers and robustness[15]. However, conventional robust DR may be subjected to a local optimal for the discontinuous and nonconvex properties of the robust M-estimator, in nonlinear DR problems[15].

In this article, a penalty function matrix is constructed with its elements representing the penalty degree of each measurement. This can be done by calculating the different types of Mahalanobis distance of observations to the main body of data and assigning small weights for outliers and large weights for other normal data. By adding penalty weights to a reconciliation object function, the influence of outliers to the estimation can be controlled, hence leading to more accurate results. The study of a continuous stirred tank reactor (CSTR) shows that the proposed robust nonlinear dynamic data reconciliation (NDDR) can reduce the effect of outliers on DR and has a better performance than the conventional NDDR method.

2 THEORY BACKGROUND

2.1 NDDR

NDDR can be formulated as a dynamic optimization problem where the objective is to minimize the deviation between the measured and the estimated values, weighted by the variance of measurement errors subjected to the dynamic model and nonlinear algebraic model and/or inequality constraints as follows[16]:

$$\min \quad \sum_{k=0}^{c}\left[\hat{\boldsymbol{y}}(t_k)-\boldsymbol{y}(t_k)\right]^{\mathrm{T}}\boldsymbol{Q}^{-1}\left[\hat{\boldsymbol{y}}(t_k)-\boldsymbol{y}(t_k)\right] \tag{1}$$

$$\text{s.t.} \quad \frac{\mathrm{d}\hat{\boldsymbol{y}}(t)}{\mathrm{d}t}=f\left[\hat{y}(t),\hat{u}(t),\hat{\theta}(t)\right]=0 \tag{2}$$

$$h\left[\hat{y}(t),\hat{u}(t),\hat{\theta}(t)\right]=0 \tag{3}$$

$$g\left[\hat{y}(t),\hat{u}(t),\hat{\theta}(t)\right]<0 \tag{4}$$

where parameter c represents the current time, k represents the sampling time, Q is the covariance matrix, f is differential equation constraints, h is algebraic equality constraints, and g is inequality constraints including simple upper and lower bounds.

NDDR presents an extra challenge for solution of the differential/algebraic optimization problem. There are two methods served in the NDDR optimization problem. One is sequential solution, wherein the nonlinear differential equations are embedded in the solution and only the initial state estimates are treated as decision variables. The differential equations are integrated using an ordinary differential equation (ODE) solver to generate the estimates for all instants within

Received 2006-09-25, accepted 2007-03-27.
* Supported by the National Natural Science Foundation of China (No.60504033).

** To whom correspondence should be addressed. E-mail: qgao@sjtu.edu.cn the time window[17]. Though straightforward, this approach is generally inefficient because it requires the accurate solution of the model equations at each iteration within the optimization even when iterates are far from the final optimal solution. The other is to use a simultaneous solution, in which the differential equations are converted to algebraic equations by discretization. These algebraic equations then solve the resulting constrained optimization problem[16]. In general, the simultaneous approach is computationally more efficient than the sequential strategy, and is used in this article.

An ideal DR scheme would use all of the information in the process measurements from the start-up of the process t_0 until the current time t_c. Unfortunately, such a scheme would necessarily result in an optimization problem of ever-increasing dimensions. For practical implementation, a moving time window[16] of length H would be used to reduce the optimization problem to manageable dimensions. If the most recent available measurements are at time step c, then a history horizon $H\Delta t$ can be defined from (t_c—$H\Delta t$) to t_c, where Δt is the time step size. So the NDDR objective function is modified as

$$\varphi=\sum_{k=c-H}^{c}\left[\hat{\boldsymbol{y}}(t_k)-\boldsymbol{y}(t_k)\right]^{\mathrm{T}}\boldsymbol{Q}^{-1}\left[\hat{\boldsymbol{y}}(t_k)-\boldsymbol{y}(t_k)\right] \tag{5}$$

2.2 M-estimator

The M-estimator is a very robust estimator. Each vector of measurement, based on its Mahalanobis distance, is assigned a weight. These weights determine the influence of each vector to the estimations so that they approach zero as the measurement error vector becomes less characteristic[18].

The most famous M-estimators used in practice include Huber estimator, Hampel's three-part redescending estimator, Tukey estimator, and Andrews estimator. All these estimators can be used to construct a penalty function matrix. In this article, Huber-type weights are used.

Huber estimator $\rho(u)$ and its influence function $\varphi(u)$ and weight factor $\omega(u)$[18]:

$$\rho(u)=\begin{cases} u^2/2, & |u|\leqslant k \\ k|u|-\dfrac{1}{2}k^2, & |u|>k \end{cases} \tag{6}$$

$$\varphi(u)=\begin{cases} u, & |u|\leqslant k \\ k\cdot\mathrm{sgn}(u), & |u|>k \end{cases} \tag{7}$$

$$\omega(u)=\begin{cases} 1, & |u|\leqslant k \\ \dfrac{k\,\mathrm{sgn}(u)}{u} & |u|>k \end{cases} \tag{8}$$

where u is the standardized residual and the parameter k controls the sensitivity of the estimator to the contaminating distribution and increases as the proportion of outliers decreases.

3 ROBUST NDDR

3.1 Robust NDDR approach

To reduce the effect of the outlier on NDDR, the objective function is modified as

$$\varphi=\sum_{k=c-H}^{c}\left[\hat{\boldsymbol{y}}(t_k)-\boldsymbol{y}(t_k)\right]^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Q}^{-1}\left[\hat{\boldsymbol{y}}(t_k)-\boldsymbol{y}(t_k)\right] \tag{9}$$

$$\boldsymbol{W}=\mathrm{Diag}\left(w_1^2,\cdots,w_j^2\right) \tag{10}$$

where $\boldsymbol{W}$ is the penalty function matrix with its elements representing the penalty degree of measurement to the estimation. The penalty function should be selected in such a way that it will decrease when measurements become increasingly less relevant to the main characteristic of the main data set. In this article, the Huber type weights are used as the penalty function.

The Huber-type weights are defined as follows:

$$w(d_i)=\begin{cases} 1, & d_i<k \\ k/d_i, & \text{otherwise} \end{cases} \tag{11}$$

where k^2 is the 90% point of χ^2 distribution with p degrees of freedom. The degrees of freedom are determined by the dimension of the observation vector. d_i is the Mahalanobis distance of the ith vector of measurements that gives its square distance from the current estimation of the mean, scaled by the current estimation of the variance.

$$d_i^2=\left[\boldsymbol{y}(t_k)-\boldsymbol{y}_{\mathrm{m}}\right]^{\mathrm{T}}\boldsymbol{Q}^{-1}\left[\boldsymbol{y}(t_k)-\boldsymbol{y}_{\mathrm{m}}\right] \tag{12}$$

The assignment of same weight to one vector will lead to loss of some valuable acceptable measurements because not all measurements of the same vector contain outliers. Therefore, once an error is detected in one vector, a further analysis is conducted within the vector. A small weight is assigned to the outlying elements with large weights being assigned to the remaining elements. To obtain element-wise weights, Mahalanobis distance d_{ij} for a single j element of the ith observation vector is given by

$$d_{ij}^2=\left[\frac{y_{ij}(t_k)-y_{\mathrm{m},j}}{\sigma_j}\right]^2 \quad i=1,\cdots,n;\ \ j=1,\cdots,p \tag{13}$$

Therefore there will be two k's, namely k_1 and k_2, corresponding to the above two different Mahalanobis distances, differing from each other in the degrees of freedom used in the calculation of k. k_1^2 is the 90% point of χ^2 distribution with p degrees of freedom and k_2^2 is the 90% point of χ^2 distribution with a freedom degree of 1.

Therefore $\boldsymbol{W}$ is constructed according to the following procedure. First, the vector of measurements is evaluated. If no error is detected, a large weight is assigned to the whole vector. Should an outlier be detected in one observation vector, an element-wise search is conducted within the vector. A small weight is assigned to the outlying elements, with large weights being assigned to the remaining elements.

In the equations cited above, $\boldsymbol{y}_{\mathrm{m}}$ and $\boldsymbol{Q}$ are location and covariance matrix of measurements in the current moving window. $y_{\mathrm{m},j}$ and σ_j are current estimations of median and variance over n observations in the current moving window.

In this article, location and variance are easily obtained from the median and median absolute deviation (MAD)[18]

$$\boldsymbol{y}_{\mathrm{m}} = \mathrm{median}\left(y\right) \tag{14}$$

$$\sigma = 1.4826\mathrm{median}\left(\left|y_i - \boldsymbol{y}_{\mathrm{m}}\right|\right) \tag{15}$$

$$\boldsymbol{Q} = \mathrm{Diag}\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\right) \tag{16}$$

where the constant 1.4826 is required to make MAD an unbiased estimation of the standard deviation for Gauss data. The advantage of MAD is that it is computationally inexpensive and relatively robust to outliers.

3.2 Correlation

If measurement errors are independent of each other, the penalty function matrix is a diagonal matrix with its diagonal elements calculated from the Huber function. For dependent error, the nondiagonal elements of the penalty function matrix are

$$w_{ij} = w_{ji} = w_i \cdot w_j \tag{17}$$

This can be proved from the definition of correlation coefficient r_{ij}

$$r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{18}$$

where σ_i and σ_j are the standard residuals of observations i and j, σ_{ij} is the covariance of observation i and j. When weighted, the variance ($\bar{\sigma}_i^2$, $\bar{\sigma}_j^2$) and covariance ($\bar{\sigma}_{ij}$) of observation i and j are

$$\bar{\sigma}_i^2 = \sigma_i^2 / w_i^2, \quad \bar{\sigma}_j^2 = \sigma_j^2 / w_j^2, \quad \bar{\sigma}_{ij} = \sigma_{ij} / w_{ij} \tag{19}$$

The correlation coefficient $\bar{r}_{ij}$ is

$$\begin{aligned} \bar{r}_{ij} &= \frac{\bar{\sigma}_{ij}}{\bar{\sigma}_i \bar{\sigma}_j} = \frac{\sigma_{ij} / w_{ij}}{\left(\sigma_i / w_i\right) \cdot \left(\sigma_j / w_j\right)} \\ &= \frac{w_{ii} w_{jj}}{w_{ij}} \cdot \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = r_{ij} \end{aligned} \tag{20}$$

From Eq.(20), it can be seen that by treatment of Eq.(17), the dependence and the original structure of the covariance matrix of measurement errors are not changed.

4 EXAMPLE

The performance of the proposed method has been tested using a simulated CSTR with a first-order exothermic reaction[16,19,20]. There are four measured variables in CSTR, two input variables: input concentration C_0, input temperature T_0, and two state variables: output concentration C, output temperature T. The process dynamic model is given by

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{q}{V}\left(C_0 - C\right) - \alpha_{\mathrm{d}} KC \tag{21}$$

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{q}{V}\left(T_0 - T\right) + \alpha_{\mathrm{d}} \frac{-\Delta H_{\mathrm{r}}}{\rho c_p} KC - \frac{-UA_{\mathrm{R}}}{\rho c_p V}\left(T - T_0\right) \tag{22}$$

where $K = K_0 \exp\left(-E_{\mathrm{A}} / RT\right)$, is an Arrhenius rate expression. The process parameters are shown in Table 1. All temperature and concentrations were scaled using a normal reference concentration ($C_{\mathrm{r}} = 1.0\,\mathrm{mol \cdot m^{-3}}$) and a normal reference temperature ($T_{\mathrm{r}} = 100\mathrm{K}$).

Measurements for all the variables were simulated at time steps of 2.5s by adding outliers and Gaussian noise to the true values, which were obtained through numerical integration of the dynamic equations. A measurement error with a standard deviation of 5% of the corresponding reference value was considered and the reconciliation of all measured variables was carried out. Outliers were added in every measured variable and the total number of outliers equalled 10% of the total data. The proposed robust NDDR algorithm was applied with a history horizon of five time steps.

The steady-state simulation was initialized at a steady-state operating point of C_0=6.5×10^6mol·m^{-3}, T_0=3.5K, C=0.1531×10^6mol·m^{-3}, and T=4.6091K. At time 30s, the feed concentration was stepped up from 6.5×10^6mol·m^{-3} to 7.5×10^6mol·m^{-3}. In this study, a simultaneous solution and optimization strategy were used to solve the dynamic optimization problem defined earlier. In this approach, the differential equations were approximated by a set of algebraic equations using the Euler discretization method. These algebraic equations were then solved with other constraints within the sequential quadratic programming (SQP) method.

The performance criteria used in this study is the mean squared error (MSE), which can be defined as

$$\mathrm{MSE}_k = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_{ik} - y_{ik}^*\right)^2 \qquad \forall k = 1, \cdots, p \tag{23}$$

where p is the number of measurement variables, N is the sampling number, $\hat{y}_{ik}$ is the estimation, and y_{ik}^* is the true value.

The estimated values for the feed concentration and feed temperature are shown in Fig.1. The estimated values for the output concentration and output temperature are shown in Fig.2. In these figures the circles correspond to the measurements, the dot line to the element-wise, distance-based estimates, the plus to the vector-wise, distance-based estimates, the solid line to the simulated (free noise) value, and the triangle to the estimates by conventional least square method. To show better detail of the dynamics, Figs.1(c), 1(d) and Figs.2(c), 2(d) are a scale of Figs.1(a), 1(b), Fig.2(a) and Fig.2(b).

Table 1 Parameters of dynamic model

q, m^3·S^{-1}	V, m^3	ΔH_{r}, J·mol^{-1}	ρ, kg·m^{-3}	c_p, J·kg^{-1}·K^{-1}	U, J·m^{-2}·s^{-1}·K^{-1}	A_{R}, m^2	T_{c}, K	K_{o}, s^{-1}	E_{A}, J·mol^{-1}	α_{d}
1.0×10^{-5}	1.0×10^{-3}	−1.13049×10^5	1.0	4187	20.935	1.0×10^{-3}	340.0	7.86×10^{12}	1.17151×10^5	1.0

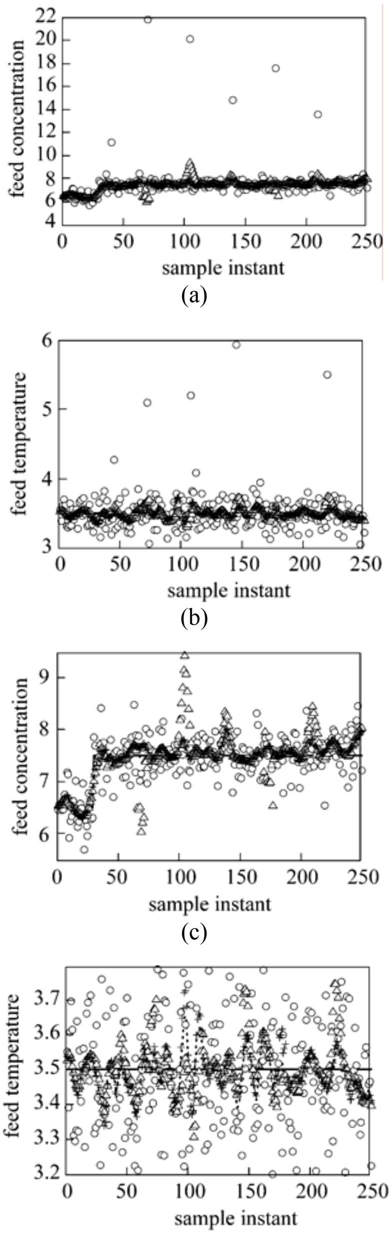

Figure 1 Feed concentration and temperature estimate's response to time step change
○ the measurements; ······ the element-wise distance-based estimates; + the vector-wise distance-based estimates; —— the simulated value; △ the estimates by conventional least square method

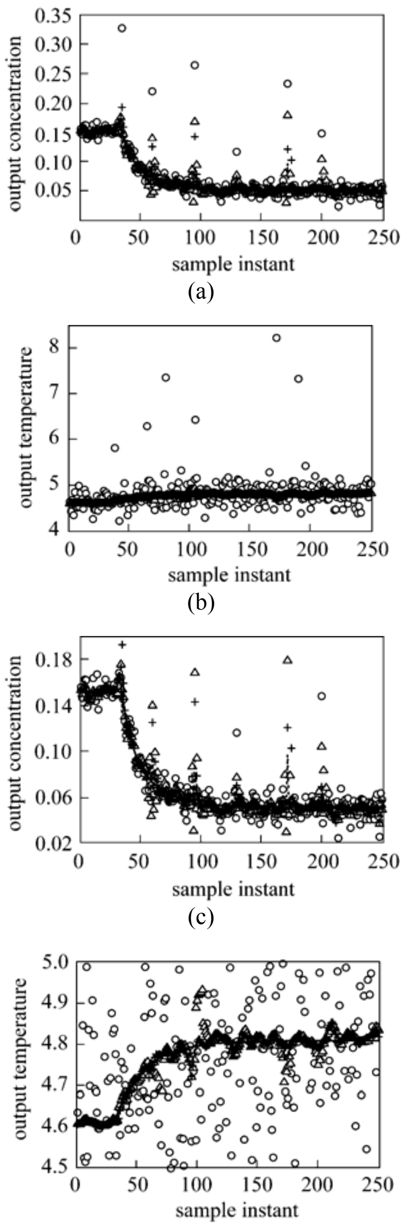

Figure 2 Output concentration and temperature estimate's response to time step change
○ the measurements; ······ the element-wise distance-based estimates; + the vector-wise distance-based estimates; —— the simulated value; △ the estimates by conventional least square method

The MSE for conventional least-square estimate, element-wise, distance-based estimate, and vector-wise, distance-based estimates are shown in Table 2.

From Figs.1 and 2 it can be seen that robust NDDR estimators give a better performance than the conventional least square method. This is because the latter tends to spread the gross error over all the measurements in the time window, leading to a biased

Table 2 Performance comparison of different estimators

Method	MSE			
	C_0	T_0	C	T
conventional least-square	0.14622	0.0056703	0.00021001	0.0008317
vector-wise distance based estimator	0.025157	0.0041709	1.0659×10^{-4}	1.9933×10^{-4}
element-wise distance based estimator	0.022195	0.0030097	3.7613×10^{-5}	1.5302×10^{-4}

estimate. When the robust NDDR is applied, small weights are assigned to the outliers, which makes the estimate insensitive to the outliers.

It can be seen from Table 2, that the element-wise Mahalanobis distance based estimator gives a better performance than the vector-wise distance based estimator. That is because the latter assigns the same penalty weight to the tagged vector, which leads to loss of information from some valuable acceptable measurements.

5 CONCLUSIONS

In this article, a robust nonlinear dynamic data reconciliation approach is proposed to reduce the influence of outliers to estimations. By introducing penalty function matrix to NDDR least square objective function, and assigning large weights to normal data and small weights to outliers, the approach reduces the influence of gross error to final estimation effectively, meanwhile maintaining the simplicity of the conventional least square estimator. Penalty weights are obtained from the weight factor of robust Huber estimator. To avoid the loss of data information, element-wise Mahalanobis distance was also proposed, as an improvement of vector-wise distance, to construct the penalty function matrix. The proposed approach can also be used in the correlative data reconciliation by suitable treatment of the nondiagonal element of the penalty matrix. An example of a dynamic CSTR verified the effectiveness of the proposed algorithm.

NOMENCLATURE

A_R	area of heat transfer, m^2
C	tank concentration, $\mathrm{mol\cdot m^{-3}}$
C_r	reference concentration, $\mathrm{mol\cdot m^{-3}}$
C_0	feed concentration, $\mathrm{mol\cdot m^{-3}}$
c	the current time, s
c_p	heat capacity, $\mathrm{J\cdot kg^{-1}\cdot K^{-1}}$
d_i	vector-wise Mahalanobis distance
d_{ij}	element-wise Mahalanobis distance
E_A	activation energy, $\mathrm{J\cdot mol^{-1}}$
f	differential equation constraints
g	inequality constraints
H	length of moving window
ΔH_r	heat of reaction, $\mathrm{J\cdot mol^{-1}}$
h	algebraic equality constraints
K	Arrhenius rate expression
K_0	rate constant, $\mathrm{s^{-1}}$
k	sampling time, s
N	sampling number
p	number of measurement variable
$\boldsymbol{Q}$	the covariance matrix
q	flow rate, $\mathrm{m^3\cdot s^{-1}}$
r_{ij}	the correlation coefficient of observation i and j
$\overline{r}_{ij}$	the correlation coefficient of observation i and j after being weighed
T	output temperature, K
T_c	coolant temperature, K
T_r	reference temperature, K
T_0	feed temperature, K
U	heat transfer coefficient, $\mathrm{kJ\cdot m^{-2}\cdot s^{-1}\cdot K^{-1}}$
V	volume, m^3
$\boldsymbol{W}$	penalty weight matrix
w_i	penalty weight factor of observation i
y_{ik}	measurement of variable k at time step i
$\hat{y}_{ik}$	estimation of variable k at time step i
$\boldsymbol{y}_\mathrm{m}$	median value of measurement vector
$\boldsymbol{y}(t)$	measurement vector at time step t
$\hat{\boldsymbol{y}}(t)$	estimated measurement vector at time step t
α_d	deactivation factor
ρ	density, $\mathrm{g\cdot m^{-3}}$
σ_i	standard deviation of variable i
σ_{ij}	the covariance of observation i and j
φ	objective function

REFERENCES

1 Kuehn, D.R., Davidson, H., "Computer control (Ⅱ): Mathematics of control", *Chemical Engineeing Progress*, **57**(6), 44—47(1961).

2 Crowe, C.M., Campos, Y.A.G., Hrymak, A., "Reconciliation of process flow rates by matrix projection (Ⅰ) Linear case", *AIChE Journal*, **29**, 881—888(1983).

3 Crowe, C.M., "Reconciliation of process flow rates by matrix projection (Ⅱ) The nonlinear case", *AIChE Journal*, **32**, 616—623(1989).

4 Sanchez, M., Romagnoli, J., "Use of orthogonal transformations in data classification-reconciliation", *Computers & Chemical Engineering*, **20**, 483—493(1996).

5 Kelly, J.D., "Formulating large-scale quantity-quality bilinear data reconciliation problems", *Computers & Chemical Engineering*, **28**(3), 357—362(2004).

6 Bagajewicz, M.J., Jiang, Q.Y., "Integral approach to plant linear dynamic reconciliation", *AIChE Journal*, **43**, 2546—2558(1997).

7 Bagajewicz, M.J., Jiang, Q.Y., "Comparison of steady state and integral dynamic data reconciliation", *Computers & Chemical Engineering*, **24**(11), 2367—2383(2000).

8 McBrayer, K.F., Edgar. T., "Bias detection and estimation in dynamic data reconciliation", *Journal of Process Control*, **5**(4), 285—289(1995).

9 Ozyurt, D.B., Pike, R.W., "Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes", *Computers & Chemical Engineering*, **28**, 381—402(2004).

10 Wang, D., Romagnoli, J.A., "Generalized T distribution and its applications to process data reconciliation and process monitoring", *Transactions of the Institute of Measurement and Control*, **27**(5), 367—390(2005).

11 Ragot, J., Chadli, M., Maquin, D., "Mass balance equilibration: A robust approach using contaminated distribution", *AIChE Journal*, **51**(5), 1569—1575(2005).

12 Tjoa, I.B., Biegler, L.T., "Simultaneously strategies for data reconciliation and gross error detection of nonlinear systems", *Computers and Chemical Engineering*, **15**(10), 679—690(1991).

13 Chen, J., Bandoni, A., Romagnoli, J.A., "Robust estimation of measurement error variance/covariance from process sampling data", *Computers & Chemical Engineering*, **21**, 593—600(1997).

14 Romagnoli, J.A., Sanchez, M.C., Data Processing and Reconciliation for Chemical Process Operations, Academic Press, London (2000).

15 Arora, N., Biegler, L.T., "Redescending estimators for data reconciliation and parameter estimation", *Computers & Chemical Engineering*, **25**(11/12), 1585—1599(2001).

16 Liebman, M.J., Edgar, T.F., Lasdon, L.S., "Efficient data reconciliation and estimation for dynamic process using non-linear programming techniques", *Computers & Chemical Engineering*, **16**, 963—986(1992).

17 Kim, I., Liebman, M.J., Edgar, T.F., "A sequential error-in-variables method for nonlinear dynamic systems", *Computers & Chemical Engineering*, **15**, 663—670(1991).

18 Huber, P.J., Robust Statistics, Wiley, New York 1989.

19 Chen, J., Romagnoli, J.A., "A strategy for simultaneous dynamic data reconciliation and outlier detection", *Computers & Chemical Engineering*, **22**, 559—562(1998).

20 Zhou, L.K., Su, H.Y., Chu, J., "A modified outlier detection method in dynamic data reconciliation", *Chin. J. Chem. Eng.*, **13**, 542—547(2005).