CHEMICAL ENGINEERING TRANSACTIONS
VOL. 74, 2019
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Sauro Pierucci, Jiří Jaromír Klemeš, Laura Piazza

ISBN 978-88-95608-71-6; **ISSN** 2283-9216
DOI: 10.3303/CET1974121

Dynamic System State Estimation and Outlier Detection Using Robust Data Reconciliation

Claudia E. Llanos[a], Mabel C. Sanchéz[a,*], Ricardo A. Maronna[b]

[a]Planta Piloto de Ing. Química-PLAPIQUI(UNS-CONICET), 8000 Bahía Blanca, Argentina
[b]Dto. de Matemática, Univ. Nac. de la Plata, 1900 La Plata, Argentina
msanchez@plapiqui.edu.ar

State estimation and detection of measurement systematic errors are critical components of plant monitoring and control procedures. Reliable estimations of the process variables are attained by Classic Dynamic Data Reconciliation procedures when measurements follow exactly a known distribution. However, if this assumption happens approximately due to the presence of systematic errors, as outliers, classic dynamic data reconciliation provides biased results. In this work, a two-step methodology of Robust Dynamic Data Reconciliation and Systematic Error Detection is proposed. It takes advantages of a moving measurement window of fixed dimension and the features of the M-estimators. Furthermore, the presence of outliers is detected using a Robust Measurement Test. Two case studies are proposed, which work with the Huber and Biweigth M-estimators. A nonlinear benchmark extracted from the literature is considered, and performance measures are reported. The results obtained demonstrate the effectiveness of the proposed methodology.

1. Introduction

The advances in measurement technology and computer data storage have allowed obtaining relevant information about the process state. With this aim different methodologies have been proposed, as the well-known Classic Data Reconciliation (CDR). This procedure minimizes the discrepancy between measurements and the process model taking advantage of the spatial redundancy provided by the model and the set of installed instruments (Romagnoli and Sánchez, 2000). Different process models have been treated with data reconciliation, this demonstrates the reliability of the strategy (Eghbal Ahmadi and Rad, 2017; Yong et al., 2018).

Classic Data Reconciliation typically assumes that measurement errors follow the normal distribution. In this case, the objective function (OF) of the optimization problem is the Least Square Estimator (LS), which provides unbiased estimates. The aforementioned assumption holds approximately due to the presence of outliers, which are atypical observations that adversely influence on the accuracy of the attained estimates. To overcome the problem, Tjoa and Biegler (1991) proposed the Robust Data Reconciliation (RDR). In this case, M-estimators are used as OF of the estimation problem because they down weight the effect of measurements with high standard error (Maronna et al., 2006). Many authors have demonstrated the improvement achieved applying RDR for steady processes (Arora and Biegler, 2001; Ozyurt and Pike, 2004; Llanos et al., 2015).

Regarding dynamic processes, the Fair, Welsch, Correntropy and Hampel functions have been used as M-estimators of the Robust Dynamic Data Reconciliation (RDDR) problem. The Fair function is a monotone M-estimator while the other three are redescendent ones. These down weight more or eliminate the effect of outliers in comparison with the Fair function. However, the application of redescendent estimators requires a good starting point of the optimization problem to avoid obtaining inaccurate solutions.

With the aim of detecting the presence of atypical measurements in RDR procedures, Albuquerque and Biegler (1996) used exploratory statistical techniques, Arora and Biegler (2001) defined explicit cutoff points obtained by minimizing the Akaike Information Criterion, Martinez Pratta et al. (2010) applied the inflection points of the M-estimator first derivative, while Zhang and Chen (2015) used the Measurement Test (MT). However, their definition of the MT is different from the one proposed by Tamhane and Mah (1982). Recently,

Paper Received: 29 April 2018; Revised: 16 October 2018; Accepted: 29 December 2018

Please cite this article as: Llanos C.E., Sanchez M.C., Maronna R.A., 2019, Dynamic System State Estimation and Outlier Detection Using Robust Data Reconciliation, Chemical Engineering Transactions, 74, 721-726 DOI:10.3303/CET1974121

a Robust MT was presented that is able to detect systematic measurement errors for linear and nonlinear steady-state processes (Llanos et al., 2017).
For implementing RDDR, the measurement moving window is updated at each time interval and then, the optimization problem is solved using deterministic (Arora and Biegler, 2001; Zhang and Chen, 2015) or stochastic (Martinez Pratta et al., 2010) procedures.
In this work, two M-estimators are selected for solving the RDDR problem using different lengths of the measurement moving window. Furthermore, the use of RMT is extended to dynamic processes whose measurements are corrupted with atypical observations. The results of comparative performance analysis are presented in terms of mean square error (MSE), Percentage of Outlier Detection (%DT) and Percentage of False alarms (%FA). The paper is structured as follows. The RDDR and RMT methodologies, and the algorithm proposed to reconcile observations and detect atypical measurements are presented in Section 2. The case studies and the performance measures are described in Section 3. The benchmark and results are included in Section 4. Finally, a Conclusions section closes this work.

2. Methodology

2.1 General Formulation of Dynamic Data Reconciliation

Let us formulate the Dynamic Data Reconciliation problem as follows:

$$
\begin{aligned}
[\hat{\mathbf{x}}_j, \hat{\mathbf{u}}_j] = \underset{x_j, u_j}{Min} \quad & \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \varphi\left(\frac{y_{ip} - x_{ij}}{\sigma_{y,i}}\right) \\
st. \quad & \mathbf{f}\left[\frac{d\mathbf{z}}{dt}, \mathbf{z}, t\right] = \mathbf{0} \\
& \mathbf{h}(\mathbf{z}, t) = \mathbf{0} \\
& \mathbf{g}(\mathbf{z}) \geq \mathbf{0}
\end{aligned}
\tag{1}
$$

where $[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R]$ are the estimated values of the measured (**x**) and unmeasured (**u**) variables at the *j*-th time interval. Furthermore φ represents the estimator function, I is the number of observations, N is the length of the measurement moving window, $\sigma_{y,i}$ is the standard deviation of the *i*-th measured variable and y_{ip} is its observed value at the *p*-th time interval. Regards the model, **z**=[**x u**], **f** and **h** represent the set of differential-algebraic process restrictions, while **g** stands for inequalities constraints.
This optimization problem can be solved using nonlinear programming techniques based on simultaneous or sequential approaches. The last one iteratively applies two steps: at first the time evolution of the process model is obtained using an ODE solver for a particular set of initial conditions, and then an optimization algorithm provides new estimations of these conditions. Both steps run until convergence for each time interval. The sequential approach has been selected for this work because it is easy to implement and a feasible solution is always obtained.

2.2 Measurement Models

Measurements always contain random errors caused by unknown and unpredictable sources. The measurement model of y_{ij} can be represented as follows:

$$y_{ij} = x_i + e_{ij} \tag{2}$$

where x_i is the true value of the i-th variable and e_{ij} stands for the random error. It is commonly assumed that $e_{ij} \sim \mathcal{N}(0, \sigma_{y,i}^2)$.
In contrast, outliers are isolated measurements that do not follow the general pattern of data causing heavy-tailed distributions. This observation is defined as follows:

$$y_{ij} = x_i + e_{ij} + K\sigma_{ij} \tag{3}$$

where K is a scalar that stands for the magnitude of the error.

2.3 Proposed strategy

Llanos et al. (2017) developed an algorithm that combines RDR and the RMT for process operating at steady state. The strategy is able to obtain reconciled measurements, detect and identify suspicious observations and classified them. In this work that algorithm is adapted to deal with dynamic processes. Next, the steps that are part of the proposed strategy are described, and the algorithm which solves dynamic problems is presented.

2.3.1 Step 1: Robust Dynamic Data Reconciliation

In this work, the Huber (HU) and Biweigth (BW) M-estimators are used as OF of the RDDR. This problem formulation for the HU estimator is the following:

$$\begin{aligned}[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R] = \underset{x_j, u_j}{Min} \quad & \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \rho_{HU}\left(\frac{y_{ip} - x_{ij}}{\sigma_{y,i}}\right) \\ st. \quad & \mathbf{f}\left[\frac{d\mathbf{z}}{dt}, \mathbf{z}, t\right] = \mathbf{0} \\ & \mathbf{h}(\mathbf{z}, t) = \mathbf{0} \\ & \mathbf{g}(\mathbf{z}) \geq \mathbf{0}\end{aligned} \tag{4}$$

where ρ_{HU} stands for the Huber M-estimator defined as

$$\rho_{HU} = \begin{cases} a^2 & |a| \leq c_{HU} \\ 2c_{HU}|a| - {c_{HU}}^2 & |a| > c_{HU} \end{cases} \tag{5}$$

and c_{HU} is a scalar. The HU function is a monotone M-estimator that only reduces the effect of atypical observations. In contrast the BW function is a redescendent M-estimator that eliminates the effect of atypical observations.

When the RDDR problem is solved using the BW M-estimators, the OF of equation (4) is replaced by the following:

$$\rho_{BW} = \begin{cases} 1 - \left[1 - (a / c_{BW})^2\right]^3 & if \quad |a| \leq c_{BW} \\ 1 & if \quad |a| > c_{BW} \end{cases} \tag{6}$$

where ρ_{BW} stands for the BW function and c_{BW} is a scalar. Due to the features of this function, a local optimal solution may be obtained. To avoid this inconvenient a good starting point is provided to the second optimization problem. This point is the solution of the one defined by Eq. (4). Both optimization problems are solved using a nonlinear programming sequential approach.

2.3.2 Step 2: Robust Measurement Test

Outliers are detected by a robust statistical hypothesis test based on the Measurement Test (Tamhane and Mah, 1982). For the j-th time interval, the RMT associates the vector of robust measurement adjustments, $\mathbf{a}_j^R$, with the robust estimate of the adjustment covariance matrix, $\widehat{\mathbf{Q}}_j^R$:

$$\mathbf{a}_j^R = \mathbf{y}_j - \hat{\mathbf{x}}_j^* \tag{7}$$

$$\hat{\mathbf{Q}}_j^R = \hat{\sigma}_a^2 \left\{ \frac{ave\left[\psi_{BW}(\mathbf{A}_j^R) / \hat{\sigma}_a\right]^2}{\left(ave\left[\psi_{BW}'(\mathbf{A}_j^R / \hat{\sigma}_a)\right]\right)^2} \right\}^T \tag{8}$$

where $\mathbf{A}_j^R$ is a matrix of dimension [I x M] that contains the last M $\mathbf{a}^R$ vectors, $\widehat{\sigma}_a$ is a scale estimate vector and $\hat{\mathbf{x}}_j^*$ is an estimation of $\hat{\mathbf{x}}_j^R$. The relation between a_{ij}^R and the i-th diagonal element of $\hat{\mathbf{Q}}_j^R$, $\hat{Q}_j^R\big|_{ii}$, gives the following statistic:

$$\hat{\tau}_{i,j}^R = \frac{\left|a_{ij}^R\right|}{\sqrt{\hat{Q}_j^R\big|_{ii}}} \sim t_{df} \tag{9}$$

which follows the Student distribution with a number of degrees of freedom, df=M-1. To provide good results the parameter M is set at values next to 30 (Llanos, thesis 2018). The level of significance of the test is set at α=0.05, and it fixes the critical statistic value tc.

2.3.3 Algorithm description

Set: maximum number of iteration S_{max}=300, tolerance Ɛ=1x10^{-8}.
Do the following procedure to obtain reliable estimates and detect the presence of outliers for the time interval [1, J-N], where J is the number of available measurement vectors:
For j=(N+1) : J

a) Read $\mathbf{x_{0j}} \in R^{Ix1}$

b) Update the MW. Define $\mathbf{Y_{ob}} \in R^{IxN}$ as the matrix which includes the last N measurements vectors

c) Define $\mathbf{Y_{ob0}} \in R^{Ix(N+1)}$ as $\mathbf{Y_{ob0}}$= [$\mathbf{x_0}$ $\mathbf{Y_{ob}}$]

d) Solve the Robust Estimation using the following iterative scheme

   i) initialize S=0 and the elements of the residue matrix $\mathbf{R_0} \in R^{Ix(N+1)}$ at 100

   ii) while S<S_{max} or max($\mathbf{R}_s$)>Ɛ

      (a) S=S+1

      (b) Solve the unrestricted optimization problem using as initial value $\mathbf{Yob_0}$. This problem solution is called $\tilde{\mathbf{X}}_{\rho,s} \in R^{Ix(N+1)}$

      (c) Solve the model with ODE 45 in the interval [j-N, j] using as starting point $\tilde{\mathbf{X}}_{\rho,s}(i,1) \forall i$. This problem solution is called $\mathbf{X}_{\text{mod,s}} \in R^{Ix(N+1)}$

      (d) Calculate the difference $\mathbf{R}_s = \left|\tilde{\mathbf{X}}_{\rho,s} - \mathbf{X}_{\text{mod},s}\right|$;

   end

   iii) Save the reconciled value: $\hat{\mathbf{x}}^R(i, j-N) = \mathbf{X}_{\text{mod}}(i,1) \ \forall \ i$

   iv) Save the new starting point $\mathbf{x}_0(i, j+1) = \mathbf{X}_{\text{mod}}(i,2) \ \forall \ i$

   v) Save the approximation of the estimated value $\mathbf{x}^*(i, j) = \mathbf{X}_{\text{mod}}(i, N+1) \ \forall \ i$

   vi) Apply the RMT $\forall \ i$ and compare each $\hat{\tau}_{i,j}^R$ with tc. If $\hat{\tau}_{i,j}^R$ > tc, an outlier is detected for the i-th variable at the j-th time interval

end

3. Performance Analysis

Two cases studies are proposed. For Case 1, the RDDR uses the HU estimator. Regarding Case 2, the BW function is applied employing as starting point the solution obtained with the HU function. One thousand simulation trials are run for each case study. Furthermore, 90% of the measurements are simulated with a standard deviation of 5% with respect to the scaled variable values, while 10% are outliers of magnitude K=10 randomly located.
The selected performance measures are the Mean Square Error (MSE), the Percentages of Detection and False Alarms (%DT -%FA) which are defined as follows

$$MSE = \frac{1}{I\ (J-N)\ SIM} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\frac{\hat{x}_{ij}^R - x_i}{\sigma_{y,i}} \right)^2, \tag{10}$$

where SIM stands for the simulation trials.

$$\%DT = \frac{Outliers\,Correctly\,Detected}{Simulated\ Out}\ 100 \tag{11}$$

$$\%FA = \frac{False\,Outliers\,Alarms}{I\ (J-N)\,SIM}\ 100 \tag{12}$$

Based on previous research, horizon lengths of magnitude 5 and 10 are considered and M is selected at 30.

4. Results

The methodology is applied to an adiabatic CSTR with a first order exothermic reactor. This nonlinear model comprises 4 states variables, two differential equations and 10 parameters, whose values are extracted from Liebman et al. (1992):

$$\frac{dA}{dt} = \frac{q}{V}(A_0 - A) - \alpha_d kA$$
$$\frac{dT}{dt} = \frac{q}{V}(T_0 - T) + \alpha_d \frac{-\Delta H_r A_r}{\vartheta C_p T_r} - \frac{UA_r}{\vartheta C_p V}(T - T_c) \quad (13)$$
$$k = k_0 \exp\left(\frac{-E_A}{T T_r}\right)$$

where k is the Arrhenious constant, A_0 and T_0 are the scaled feed concentration and temperature, A and T stand for the scale tank concentration and temperature, respectively, and A_r and T_r are nominal reference values. The CSTR simulations are started at A_0=6.5 and T_0=3.5. After 30 times, a step change of magnitude 1 is introduced to A_0.

Figure 1 and 2 show the set of measurements of one simulation. Also, they display the estimated A and T values obtained using the LS, HU function (Case 1) and the BW initialized with the solution attained using HU (Case 2). It is observed a reduction of variance since the measurements present more dispersion than the reconciled variables represented by curves. It can be seen that better results are achieved for Case 2 and both N values.

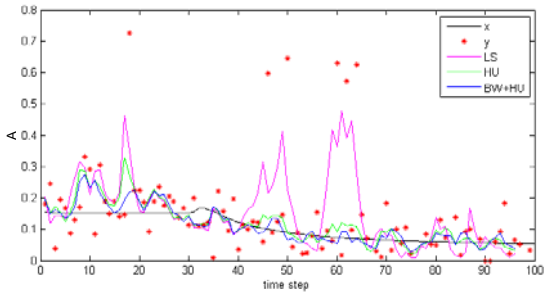

Figure 1a: Estimation of concentration for N=5

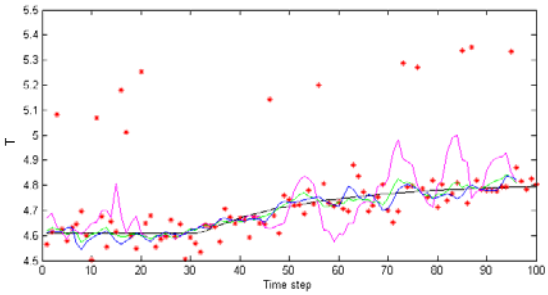

Figure 1b: Estimation of Temperature for N=5

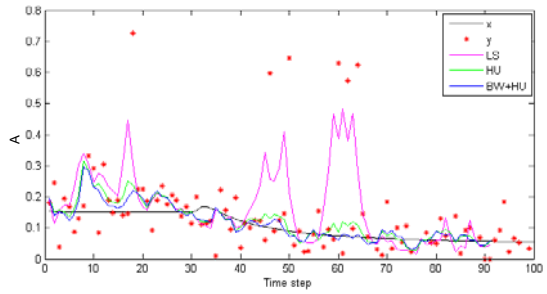

Figure 2a: Estimation of concentration for N=10

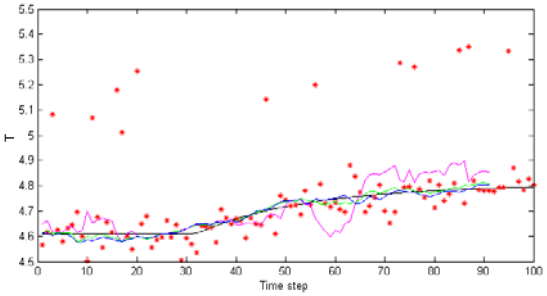

Figure 2b: Estimation of Temperature for N=10

The overall results of the simulation trials are presented in Tables 1 and 2. Table 1 shows the accuracy of the estimations in terms of the MSE. Better results are achieved with RDDR for both values of N analyzed. The best MSE is attained for Case 2 and the longest N; this result is according to the theory of robust estimators.

Table 1: MSE

N	LS	Case 1	Case 2
5	4.15693	0.70392	0.59471
10	3.74236	0.51425	0.34811

Table 2 shows that similar %DT and %FA are achieved with both case studies. These measures decrease as N increases, which indicates that a bigger fraction of outliers is detected with N=5 although the reconciled value is more accurate when N=10. This occurs because the RMT uses an approximation of the reconciled value, which is not accurate, for its statistic computation. Better results are obtained if the RMT is applied to the measurements that correspond to the reconciled value, however errors are detected N times latter they happen. A compromise between the number of detected outliers and the detection time is observed.

Table 2: Percentage of Detection and False Alarms

	M=30			
	HU		BW + HU	
N	% DT	%FA	%DT	%FA
5	94.581	2.651	94.556	2.557
10	89.655	2.410	89.609	2.398

5. Conclusions

The comparison between RDDR and CDD shows the improvement of accuracy obtained with the robust methodology. The MSE diminishes when N increases because more accurate results are achieved. The best results are attained using the Biweigth estimator and N=10.

Regards the RMT, similar %DT and %FA are achieved for a given N. In both case studies, more than 89% of outliers simulated are correctly detected, while false alarms are lower than 2.7%. Furthermore, better results are achieved when N=5 because the RMT uses an approximation of the reconciled value that is more accurate for this window length.

The performance measures indicate that the RDDR and the RMT are methodologies that can be applied for detecting outliers in dynamic systems. It is observed a trade-off between the accuracy achieved using the RDDR and the detection capability. For the simulated case study best results are obtained using BW as OF.

Acknowledgments

The authors acknowledge the financial support of CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) and UNS (Universidad Nacional del Sur, Bahía Blanca, Argentina).

References

Albuquerque, J.S., Biegler, L.T., 1996. Data reconciliation and gross-error detection for dynamic systems. AIChE journal, 42(10), 2841-2856.

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

Eghbal Ahmadi M.H., Rad A., 2017, Plant-wide simulation model for modified claus process based on simultaneous data reconciliation and parameter estimation, Chemical Engineering Transactions, 57, 997-1002

Leibman, M. J., Edgar, T. Lasdon, L.S., 1992. Efficient data reconciliation and estimation for dynamic processes using nonlinear programming techniques. Computers & chemical engineering, 16(10-11), 963-986.

Llanos, C., Sánchez, M., R. Maronna, 2015, Robust estimators for data reconciliation, Industrial. & Engineering. Chemistry Research, 54, 5096-5105.

Llanos, C., Sánchez, M., R. Maronna, 2017, Classification of systematic measurement errors within the framework of robust data reconciliation, Industrial & Engineering Chemistry Research, 56, 9617-9628.

Llanos, C., 2018, Metodologías robustas de reconciliación de datos y tratamiento de errores sistemáticos, PhD Thesis, Universidad Nacional del Sur. Dpto de Ingeniería Química, Argentina

Mah, R. S. H.; Tamhane, A. C., 1982, Detection of gross errors in process data. AIChE journal, 28, 828-830.

Maronna, R. A.; Martin, R. D. and V. Yohai (Ed.1), 2006, Robust Statistics: Theory and Methods, John Wiley and Sons Ltd.: Chichester, UK.

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

Prata, D. M., Schwaab, M., Lima, E. L., Pinto, J. C., 2010, Simultaneous robust data reconciliation and gross error detection through particle swarm optimization for an industrial polypropylene reactor. Chemical Engineering Science, 65(17), 4943-4954.

Romagnoli, J., and Sánchez, M. (Ed), 2000, Data Processing and Reconciliation for Chemical Process Operations; Academic Press: San Diego, USA.

Tjoa, I. B., Biegler, L. T., 1991, Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems. Computers & Chemical Engineering, 15, 679–690.

Zhang, Z. and J. Chen, 2015, Correntropy based data reconciliation and gross error detection and identification for nonlinear dynamic processes, Computers & Chemical Engineering, 75, 120-134.

Yong J.Y., Varbanov P.S., Klemes J.J., 2018, Data reconciliation focusing on the utility system of a total site , Chemical Engineering Transactions, 70, 1987-1992.