Antonio Espuña, Moisès Graells and Luis Puigjaner (Editors), Proceedings of the 27th European Symposium on Computer Aided Process Engineering – ESCAPE 27
October 1st - 5th, 2017, Barcelona, Spain © 2017 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/B978-0-444-63965-3.50256-7

A robust methodology for the sensor fault detection and classification of systematic observation errors

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

[a]*PLAPIQUI, Departamento de Ingeniería Química, Universidad Nacional del Sur (UNS),CONICET, Camino La Carrindanga km 7, Bahía Blanca 8000, Argentina*

[b]*Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata 1900, Argentina*
msanchez@plapiqui.edu.ar

Abstract

Robust Data Reconciliation enhances the quality of variable estimates when the data set contains a moderate proportion of atypical observations. But if systematic errors that persist in time, i.e. biases and drifts, are not detected, the break down point of the estimates is exceeded and results get worse. In this work, a new methodology based on the concepts of Robust Statistics is presented to deal with this problem. The strategy computes robust variable estimates, classifies the systematic measurement errors, and provides corrective actions to avoid the detrimental effect of biases and drifts until the sensor is repaired. The performance of the methodology is evaluated for the steady state operation of linear and non-linear benchmarks. Results demonstrate that its use significantly improves the estimates accuracy.

Keywords: Data Reconciliation, Robust Statistics, Measurement Errors.

1. Introduction

Many researchers have proved the advantages of using robust M-estimators for the resolution of the Data Reconciliation (DR) problem under the presence of outliers (Llanos et al., 2015). Also the performance of those estimators to reduce the effect of measurements contaminated with systematic errors that persist in time (SEPT), as biases (Chen et al., 2013) and drifts (Nicholson et al., 2014), has been analysed.

In contrast, the robust classification of systematic errors has been addressed in few published works. This classification helps to identify faulty instruments before the Break Down Point (BDP) of the estimates is reached, take corrective actions to reduce the Mean Square Error (MSE) of the estimations, and provide information to the plant instrumentation sector that allows a quick repair of the fault. Regarding this topic, at first the simultaneous identification of biases and outliers using the Welsch estimator was considered by Martinez Prata et al. (2013). Recently the presence of outliers, biases and drifts was distinguished applying the Correntropy estimator (Zhang et al., 2015). Performance evaluation studies of these methodologies were not provided.

In this work, a new strategy for the robust classification of systematic measurement errors is presented. The DR stage is based on the Simple Method, SIM, (Llanos et al., 2015). A Robust Measurement Test (RMT) is formulated to detect suspicious variables and identify outliers. Instead, biases and drifts are classified using a Robust Linear Regression (RLR). Application results of the methodology for a plant water distribution

system, PWD, and a subsystem of the Tennessee Eastman Process, TENs, are used to evaluate its performance. This is measured in terms of the MSE, the Percentage of SEPT Detection, and the Percentage of Right Identification for outliers, biases and drifts.

2. Measurement Models

Random errors are caused by unknown and unpredictable changes in the instrument or in the environmental conditions, and they often follow a Gaussian distribution. These errors generate inconsistencies between the measurements and the plant balance equations, which can be reduced applying the classical DR procedure. The presence of systematic measurement errors causes deviations with respect to the normal distribution. In this case, Robust DR methodologies are applied because they provide more reliable estimates than the classical one in the presence of a moderate amount of gross errors.

Outliers are isolated errors whose detrimental effect on variable estimates can be reduced by applying robust DR. In contrast, biases and drifts are SEPT which can affect variable estimates if their BDPs are exceeded. The measurement models (Eq.1 to Eq.4) considered in this work are presented in Table 1, where the indexes i and j are used to indicate the variable and the time interval, respectively.

Table 1. Measurement Models

Error Measurement	Model	Eq.	Nomenclature
Random	$y_{ij} = x_i + e_{ij}$	(1)	x_i: true value of the variable y_{ij}: measurement value e_{ij} :random error
Outlier	$y_{ij} = x_i + e_{ij} + O_{ij}$	(2)	$O_{ij} = K^o_{ij}\,\sigma_i$ σ_i : standard deviation K^o_{ij}: outlier magnitude
Bias	$y_{ij} = x_i + e_{ij} + B_i(t)$	(3)	$B_i(t) = K_i^b\,\sigma_i$ K_i^b:bias magnitude
Drift	$y_{ij} = x_i + m_{drift} f(t)$	(4)	m_{drift}: constant f(t): function of time

3. New Robust Data Reconciliation Strategy

The proposed strategy involves three methodologies (robust DR, RMT, and RLR) which work together. Next, they are briefly described.

3.1 Robust Data Reconciliation

The SIM provides good results considering both estimates accuracy and computational requirements. This methodology has two sequential steps which take advantages of the temporal redundancy and the features of the monotone and redescending M-estimators.

Step 1- At the j-th time interval, the robust median of the i-th variable, $\hat{y}_{ij}^R$ (i=1,…,I), is calculated using the data included in a moving window of length N $\{y_{ip}, p=j-N+1,...,j\}$:

$$\hat{y}_{ij}^{R} = \underset{y_{ij}}{Min} \sum_{p=j-N+1}^{j} \rho_{BW}\left(\frac{y_{ip} - y_{ij}}{\sigma_i}\right) \tag{5}$$

where ρ_{BW} is the Biweigth estimator.

Step 2- Using the solution of Step 1 as starting point, an optimization problem is solved to estimate the vectors ($\hat{x}_j^{SIM}$,$\hat{u}_j^{SIM}$) that represent the state of the system at time interval j

$$[\hat{x}_j^{SIM} \quad \hat{u}_j^{SIM}] = \underset{x_j}{Min} \sum_{i=1}^{I} \rho_H\left(\frac{\hat{y}_{ij}^{R} - x_{ij}}{\sigma_i}\right)$$
$$s.t. \tag{6}$$
$$f(x_j, u_j) = 0$$

where ρ_H is the Huber estimator and u_j is the unmeasured variable vector.

3.2. Robust Measurement Test

At the time interval j, the vector of measurement adjustments $a_j^R = y_j - \hat{x}_j^{SIM} \sim N(0, Q^R)$ if only random errors are present. To decide if the error of measurement $y_{i,j}$ is random or not, a statistical hypothesis test based on the concepts of Robust Statistics is developed as extension of the Measurement Test (Narasimhan and Jordache, 2000). The proposed test statistic is

$$t_{i,j}^{R} = \frac{\left|a_{i,j}^{R}\right|}{\sqrt{Q_{ii}^{R}}} \tag{7}$$

where $a_{i,j}^R = y_{i,j} - \hat{x}_{i,j}^{SIM}$ is the robust measurement adjustment of the i-th variable and Q_{ii}^R is the i-th diagonal element of Q^R. Because Q^R is unknown, a robust estimation of this matrix is calculated at time j, $\hat{Q}_j^R$. With this purpose, the matrix A_j^R is formed containing the last a_p^Rvectors (p=j-N+1,…,j), and $\hat{Q}_j^R$ is evaluated as follows

$$\hat{Q}_j^R = \hat{\sigma}_a^2 \left\{ \frac{\text{ave}\left[\psi(A_j^R)/\hat{\sigma}_a\right]^2}{\left(\text{ave}\left[\psi'(A_j^R/\hat{\sigma}_a)\right]\right)^2} \right\}^T \tag{8}$$

where ψ is the derivative of ρ_{BW}, $\hat{\sigma}_a^2$ is an scale estimate vector of each row of A_j^R, that is calculated as the square of the normalized median absolute deviation about the median (MADN) and ave represents the sample average. Using $\hat{Q}_j^R$ in Eq.(6) the statistic is reformulated:

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

This follows the Student distribution with a number of degree of freedom, df=N-1. The level of significance of the test is set at α=0.025. To fix the critical values t_{ck} (k=1,...,4), the probability of occurrence of consecutives errors is considered.

3.3 Robust Linear Regression

Robust regression methods aim at giving a good fit to the bulk of the data without being perturbed by a small proportion of gross measurement errors. A linear regression model for any variable is represented by Eq.(9). It is fitted to a data set $\{(x_m, y_m): m=1,\ldots,M\}$, where x_m and y_m are the predictor and response variable values and M is the total quantity of measurements considered for the regression.

$$y = \beta_0 + \beta_1 x \tag{10}$$

To calculate the vector $\hat{\beta}$, the following optimization problem is solved:

$$Min \sum_{m=1}^{M} \rho_{BW}\left(\frac{r_m(\hat{\beta})}{\hat{\sigma}_r}\right) \tag{11}$$

where $r_m=y_m-(\hat{\beta}_0+\hat{\beta}_1 x_m)$, and the scale estimation $\hat{\sigma}_r$ is calculated as the MADN. The necessary and sufficient condition for solving the problem formulated in Eq.(11) is:

$$\sum_{m=1}^{M} \rho'_{BW}\left(\frac{\hat{r}_m}{\hat{\sigma}_r}\right) X_m = 0 \tag{12}$$

To distinguish between a bias or a drift, the following statistical hypotheses are confronted: H_0: $\beta_1=0$ and H_1: $\beta_1 \neq 0$. If the first one is not rejected, it is considered that the measurement is contaminated with a bias. To take a decision between both hypotheses, the statistical hypothesis test $T_{\beta 1}$ is formulated as the relation between $\hat{\beta}_1^R$ and its variance. This statistic follows the Student Distribution with df=M-2:

$$T_{\beta_1} = \frac{\hat{\beta}_1^R}{\left[\text{var}\left(\hat{\beta}_1^R\right)\right]^{1/2}} \sim t_{df} \tag{13}$$

If $T_{\beta 1}$ is lower than its critical value for $\alpha=0.05$, the SEPT is classified as a bias and the parameter $\hat{\beta}_0$ represents its magnitude. In contrast the error is considered as a drift.

3.4 New Robust Data Reconciliation Methodology

For each time interval, the vector y_j is stored into the window data matrix Y and the oldest measurement vector is eliminated. The suspicious variables detected in previous times, which were saved in vector s, are inspected. If s is empty the DR is run using Y, in contrast the measurements of the variables contained in s are replaced by appropriate values.

The statistic of the RMT is calculated for each variable and compared with t_{c1}. If $\hat{t}_{i,j}^R$ is greater than t_{c1}, an atypical observation is detected for the i-th variable. If its next statistic is lower than t_{c2} the previous measurement is classified as an outlier, else the variable continues been analyzed. Once four consecutive statistic values are greater than their respective t_{ck}, it is considered a suspicious variable. In contrast, consecutives outliers are identified.

A suspicious variable is saved in s, and the arrival of new measurements is waited before the estimation of the linear regression model. Once N/2 observations are collected, the evaluation of $T_{\beta 1}$ follows to classify the SEPT. The information provided by the classification is sent to the DR stage. Measurements with drift are replaced by others generated using $\hat{x}_{j-4}^{SIM}$, while measurements contaminated with biases are corrected using the bias magnitude. This is calculated as the difference between the robust median of the measurements contained in Y and the reconciled value $\hat{x}_j^{SIM}$.

4. Performance Analysis

Two benchmarks are used to demonstrate the methodology performance. For each benchmark, two case studies are proposed. The values of N are changed for fixed magnitudes of outliers (K_i^o=10), biases (B_i =6) and drifts (m_{drift}=1). Case 1 shows the behaviour of the robust DR procedure when measurements are corrupted with SEPT, while Case 2 remarks the advantages of using the proposed strategy for the same conditions.

Five thousand simulation trials are run for each case study. The probability of gross errors is fixed at 0.02, these are randomly generated. Once a SEPT is developed, it persists during 100 time intervals; therefore 10 % of the simulated measurements are contaminated with atypical values. The selected performance parameters are the MSE, the percentage of SEPT detection (%DTs) and the correct and wrong identification of outliers (% OI), biases (% BCI-% BWI) and drifts (% DCI-% DWI).

$$\%DTs = \frac{SEPT\ Correctly\ Identified}{Simulated\ SEPT}\ 100 \tag{14}$$

$$\%XCI = \frac{X\ Correctly\ Identified}{Simulated\ X}\ 100, \quad X \in [O, B, D] \tag{15}$$

$$\%XWI = \frac{X\ Detected\ and\ Wrongly\ Identified}{Simulated\ X}\ 100, \quad X \in [B, D] \tag{16}$$

5. Results

The analysed benchmarks are the PWD, which is composed by 82 variables and 47 linear balance equations, and the feed flows and the reactor of TENs that involves 44 variables and 16 component balances with reaction. Measurement standard deviations are 2.5% and 2% of the true values, respectively. The results of the simulation trials are presented in Tables 2 and 3 for the aforementioned performance indexes.

The PWD gets its best performance when N=30 and the %DTs is the highest. When N increases the MSE grows as the %DTs decreases and the base case MSE also grows.

For TENs, the best MSE is obtained when the %DWI=0, although the %DTs is not the best one. This shows that detection is as important as the correct identification of SEPT.

The comparison of Case 1 and 2 highlights the reduction on the MSE achieved with the developed methodology. For the analysed windows length, the % DTs is always higher than 92 %. For the PWD, sensors with bias can work on-line for the 96 % of the faults.

Table 2. PWD performance indexes

N	Case 1	Case 2						
	MSE	MSE	% DTs	% OI	BCI %	DCI %	% BWI	% DWI
20	342.688	14.957	92.33	96.70	96.70	93.05	3.14	0.00
30	393.566	1.135	96.73	96.18	96.18	97.37	5.77	0.00
40	438.229	28.626	96.31	96.59	96.59	96.53	5.33	0.00

Table 3. TENs performance indexes

N	Case 1	Case 2						
	MSE	MSE	% DTs	% OI	BCI %	DCI %	% BWI	% DWI
20	121.063	5.931	95.38	92.9	91.67	95.51	1.19	2.25
30	102.519	5.467	97.66	94.82	88.57	98.17	7.62	0.92
40	105.843	2.34	94.81	95.76	85.32	95.15	9.17	0.00

6. Conclusions

The analysis of performance global measures indicates that the MSE diminishes because N grows or a best percentage of detection is achieved.

The analysis of individual indexes demonstrates the importance of the right identification of SEPT. Their wrong identification affects the MSE, especially when they are drifts. That is because an erroneous corrective action is sent to the DR problem.

Global and individual indexes are useful to show the performance of the RMT and RLR procedures, respectively. The MSE increases both when errors are not detected or well classified. Thus, both the detection and the right identification are essential to increase the quality of variable estimates.

References

J. Chen, Y. Peng, and J. Munoz, 2013, Correntropy Estimator for Data Reconciliation, Chemical Engineering Science, 104, 1019-1027.

C. E. Llanos, M. C. Sánchez, and R. A. Maronna, 2015, Robust Estimators for Data Reconciliation. Industrial and Engineering Chemistry Research, 54, (18), 5096–5105.

D. Martinez Prata, M. Schwaab, E.L. Lima, and J.C. Pinto, 2013, Simultaneous Robust Data Reconciliation and Gross Error Detection through Particle Swarm Optimization for an Industrial Polypropylene Reactor, Chemical Engineering Science, 65, (17), 4943-4954.

S. Narasimhan, Jordache, 2000, Data Reconciliation and Gross Error Detection,Gulf Publishing Company, United States of America.

B. Nicholson, R. López-Negrete, and L.T. Biegler, 2014, On-line State Estimation of Nonlinear Dynamic Systems with Gross Errors, Computers and Chemical Engineering, 70, 149-159.

Z. Zhang, and J. Chen, 2015, Correntropy Based Data Reconciliation and Gross Error Detection and Identification for Nonlinear Dynamic Processes,Computers and Chemical Engineering, 75, 120-134.