I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

Robust Estimators for Data Reconciliation

Claudia E. Llanos,[†] Mabel C. Sanchéz,*[,†] and Ricardo A. Maronna[‡]

[†]Planta Piloto de Ingeniería Química, Universidad Nacional del Sur−CONICET, Camino La Carrindanga km 7, 8000 Bahía Blanca, Argentina

[‡]Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Avenida 7 776, 1900 La Plata, Argentina

ABSTRACT: In this work, a comparative performance analysis of robust data reconciliation strategies is presented. The study involves two procedures based on the biweight function and three estimation techniques that use the Welsh, quasi-weighted least squares, and correntropy M-estimators. The aforementioned functions are selected for comparative purposes because their use in the data reconciliation literature has appeared during the past decade. All procedures are properly tuned to have the same estimation and gross error detection/identification capabilities under the ideal distribution. Different measurement models are systematically taken into account, and results are analyzed considering both performance measures (average number of type I errors, global performance, mean square error) and computational load. The comparative analysis indicates that a simple robust methodology can provide a good balance between those two issues for linear and nonlinear benchmarks.

1. INTRODUCTION

The operation of today's chemical plants is characterized by the stringent need of introducing fast and low-cost changes to improve their performance. The decision-making process about possible modifications in a system requires knowing its actual state. This is determined by the values of the process variables contained in the model chosen to represent plant operation. In general, this model is constituted by the mass and energy conservation equations.

During the normal operation of a chemical process, measurements such as flow rates, temperatures, pressures, compositions, etc. are obtained. The numerical values resulting from the observations do not provide consistent information because they contain some type of error that prevents the conservation equations from being exactly satisfied. Therefore, it is a common practice to apply data reconciliation procedures that provide adjusted measurements values, which are consistent with the corresponding balance equations.[1,2]

Different approaches have been proposed for the simultaneous treatment of random errors and outliers in data reconciliation problems. Instead of minimizing the least-squares (LS) criteria, which is strongly biased by the presence of outliers, Tjoa and Biegler[3] initially formulated an objective function based on the contaminated normal distribution following the maximum likelihood principle. When the procedure converges, an observation is identified as an outlier if its contribution to the sample probability is greater than the corresponding to the random error. The performance of this method strongly depends on an adequate characterization of the gross error. Furthermore, it often leads to nonconvex and complex objective functions that are prone to underflow problems.[4]

Other simultaneous approaches based on the concepts of robust statistics have been devised since that time. Robust strategies produce reliable estimates not only when data follow a given distribution exactly, but also when this happens only approximately due to the presence of outliers.[5,6]

Different types of M-estimators, which are generalizations of the maximum likelihood estimator, have been used as objective functions of the data reconciliation problem instead of the weighted LS. Albuquerque and Biegler[4] employed a convex M-estimator, the fair function (FF), which has the interesting property of yielding global optima for nonlinear problems with low constraints curvature. Since this estimator provides no direct inference to detect outliers, techniques based on exploratory statistics were used for that purpose. Next, Arora and Biegler[7] applied the three-part redescending estimator (TPRE) proposed by Hampel,[8] which has superior robustness with respect to the FF. In this case, outlier detection and identification can be performed using an explicit cutoff point. The parameters of the redescending estimator were tuned for a specific application by minimizing the Akaike information criterion. The estimation obtained using the FF was used as the starting point for minimizing the TPRE.

A partially adaptive estimator based on the generalized T-distribution and a fully adaptive estimator based on density estimation were proposed by Wang and Romagnoli.[9] These methods showed improved robustness and efficiency in comparison to traditional approaches at the expense of increasing the computational load.

During the past decade, the contribution by Ozyurt and Pike[10] has been highlighted. The authors presented a performance analysis of seven objective functions, which have been previously used to solve data reconciliation problems, and three gross error detection criteria. Both simulated and industrial processes operating at steady state were considered. For comparison purposes, the objective functions were tuned to obtain the same relative efficiency at the ideal distribution. Promising results were attained using the Cauchy distribution and the TPRE.

Regarding the application of robust statistics to address dynamic data reconciliation problems, Prata et al.[11] performed a comparative analysis that involved the Welsch (W) M-estimator

Received: December 3, 2014
Revised: April 6, 2015
Accepted: April 7, 2015

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

and the same objective functions studied by Ozyurt and Pike.[10] The Lorentzian and W distributions provided the best reconciled values for the analyzed case studies. Later on, the W function was used to adjust the measurements and estimate the parameters of an industrial polypropylene reactor[12] using a moving window approach.

The strengths of monotone and redescending M-estimators were combined by Sánchez and Maronna[13] who presented two strategies based on the Huber (H, monotone) and biweight (BW, redescending) M-estimators. Computing was executed using the regression approach proposed by Maronna and Arcas.[14] Simulated measurements for the steam metering network (SMN) benchmark[15] were reconciled taking into account different observations models. Results were compared with those provided by the TPRE. The relative efficiency and the average number of type I errors (AVTI) at the ideal distribution were set at the same values for all the strategies. A slightly superior behavior of the proposed methodologies with respect to TPRE was achieved with lower computational time demand.

The quasi-weighted least square (QWLS) M-estimator was formulated by Zhang et al.[16] for data reconciliation. The Akaike information criterion was used to tune the estimator parameter for each specific application, and the cutoff point was set by selecting the probability of committing type I errors at the ideal distribution. Estimator performance was compared with respect to the behavior of the FF and TRPE for the SMN benchmark, and authors pointed out that the QWLS estimator was more effective than the other ones.

Later on, Chen et al.[17] proposed to use correntropy (CO) as an optimality criterion in estimation problems. The effectiveness of the CO estimator was tuned by minimizing the Akaike information criterion. The cutoff point was also determined by selecting the probability of committing a type I error at the ideal distribution. A performance comparison study was performed for the SMN benchmark, and it was concluded that CO function outperformed the QWLS estimator for that case study.

Recently, dynamic data reconciliation problems were addressed using the FF, the TPRE,[18] and the CO function[19] taking into account sensor drifts and biases.

In this work, the robust strategies that have appeared in the data reconciliation literature during the past decade are reviewed, and the results of a comparative performance analysis accomplished for two benchmarks are presented. Different measurement models are systematically taken into account, and results are provided in terms of the performance measures usually evaluated for this type of studies: AVTI and Global Performance (OP), proposed by Narasimhan and Mah,[20] and the mean square error (MSE). Under the ideal distribution, the estimation and outlier detection/identification capabilities of all the strategies are the same. This is guaranteed by an adequate parameter tuning.

The paper is structured as follows. In Section 2, the robust strategies selected for this study are briefly reviewed. The procedure devoted to evaluate the performance of the techniques is described in Section 3. The results of the comparative analysis are presented in Section 4, and a Conclusions section closes the article.

2. ROBUST ESTIMATORS FOR DATA RECONCILIATION

Let the state of a plant operating under steady-state conditions be described by the vectors $\mathbf{x} = (x_1,..., x_n)$ and $\mathbf{u} = (u_1,..., u_m)$ of measured and unmeasured process variables, respectively, which satisfy a system of independent nonlinear balance equations:

$$f(\mathbf{x}, \mathbf{u}) = 0 \tag{1}$$

In the absence of systematic errors, consider the following measurement model

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

where y_{ij} represents the measurement of the ith variable at the jth time period, and e_{ij} stands for the unobservable independent random error, which has zero mean and known standard deviation σ_i. Call $\mathbf{y}_j = (y_{1j},..., y_{nj})$ the vector of observations at time interval j.

An M-estimator of the state of the system at time t is defined as the solution $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$, satisfying eq 1, of

$$\sum_{j=t-N+1}^{t} \sum_{i=1}^{n} \rho\left(\frac{y_{ij} - x_i}{\sigma_i}\right) = \min \tag{3}$$

where

$$r_{ij} = \frac{y_{ij} - x_i}{\sigma_i} \tag{4}$$

stands for the standardized residual, and ρ is the estimator's loss function. It is an increasing function of $|r|$, and the case $\rho(r) = r^2$ corresponds to the LS estimator. The process is assumed to be observed for a data horizon of size N, that is, at time t, the estimator is based on observation vectors $\mathbf{y}_{t-N+1},..., \mathbf{y}_t$. Eq 3 can be solved for a single observation, but the combined use of temporal and spatial redundancy provides better estimates.

The W, BW, QWLS, and CO functions have appeared as loss functions to solve the robust data reconciliation problem during the past decade. Those are displayed in Figure 1 for a relative efficiency of 95% at the standardized normal distribution. Let the derivative of ρ be $\psi = \rho'$ and the weight function W, which is useful to express a location M-estimate as a weighted mean,[6] be defined as follows:

$$W(r) = \begin{cases} \psi(r)/r & \text{si } r \neq 0 \\ \psi'(0) & \text{si } r = 0 \end{cases} \tag{5}$$

Both ψ and W functions are represented in Figures 2 and 3, respectively.

Next, robust data reconciliation strategies based on the above-mentioned loss functions are briefly reviewed. First, the simple (SiM) and sophisticated (SoM) methods, which make use of the BW M-estimator, are presented. Then, the W, QWLS, and CO M-estimators are introduced.

2.1. Simple and Sophisticated Methods. These are nonadaptive techniques that combine the strength of monotone and redescending M-estimators.[13] The SiM involves the following two steps.

2.1.1. Step 1. A location M-estimate from the BW family is computed for the ith measurement $(i = 1{:}n)$ at time t, $\tilde{y}_{it}$, which is the solution of

$$\sum_{j=t-N+1}^{t} \rho_{\mathrm{BW}}\left(\frac{y_{ij} - \tilde{y}_i}{\sigma_i}\right) = \min \tag{6}$$

where

$$\rho_{\mathrm{BW}} = \begin{cases} 1 - \left[1 - (r/c_{\mathrm{BW}})^2\right]^3 & \text{if } |r| \leq c_{\mathrm{BW}} \\ 1 & \text{if } |r| > c_{\mathrm{BW}} \end{cases} \tag{7}$$

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

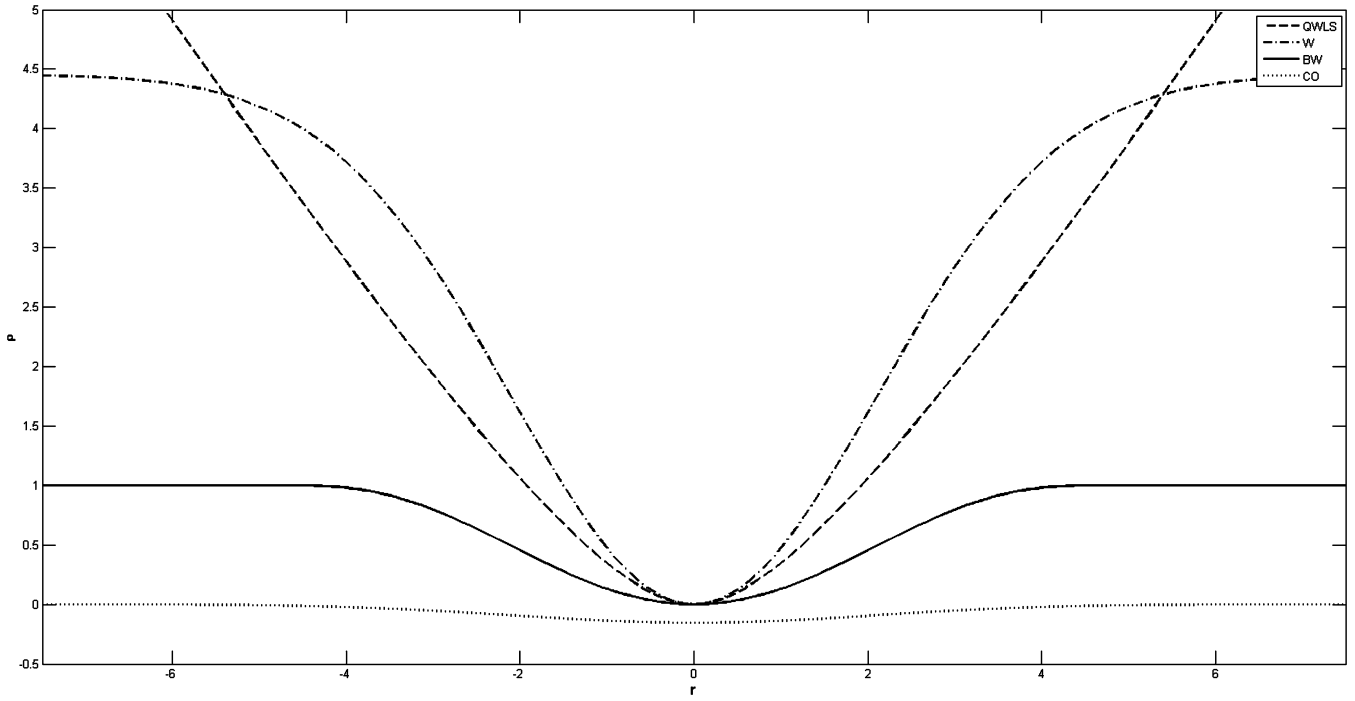

Figure 1. Loss functions.

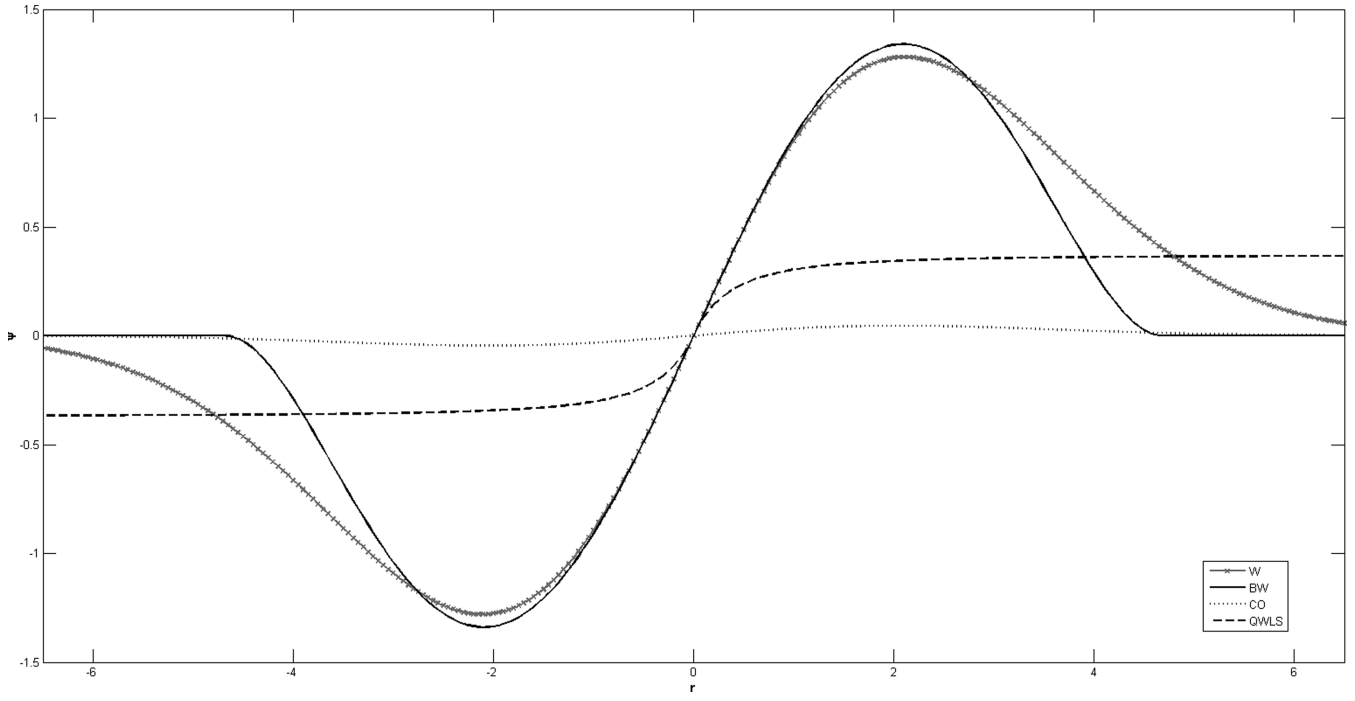

Figure 2. Influence functions.

The estimator $\tilde{y}_{it}$ is based on the time horizon values $\{y_{ij}, j = t - N + 1,...,t\}$. In this way, a robust and simple initial estimator, namely the median of $\{y_{t-N+1},..., y_t\}$, is available for each measurement. Since for the estimation of x_i the use of redundancy from observations other than i is limited,[14] full advantage of the redundancy supplied by the repeated observations y_{ij} in the time horizon is taken. The location estimator uses this "self-redundancy" to detect and down-weight outliers.

2.1.2. Step 2. An M-estimator of the process state at time t is defined as the solution $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$, satisfying eq 1, of

$$\sum_{i=1}^{n} \rho_{\mathrm{H}}\left(\frac{\tilde{y}_i - x_i}{\sigma_i}\right) = \min \tag{8}$$

where ρ_{H} corresponds to the Huber family

$$\rho_{\mathrm{H}} = \begin{cases} r^2 & \text{if } |r| \le c_{\mathrm{H}} \\ 2c_{\mathrm{H}}|r| - {c_{\mathrm{H}}}^2 & \text{if } |r| > c_{\mathrm{H}} \end{cases} \tag{9}$$

These two steps work together as follows. The first one helps to down-weight the effect of outliers that may bias the estimation

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

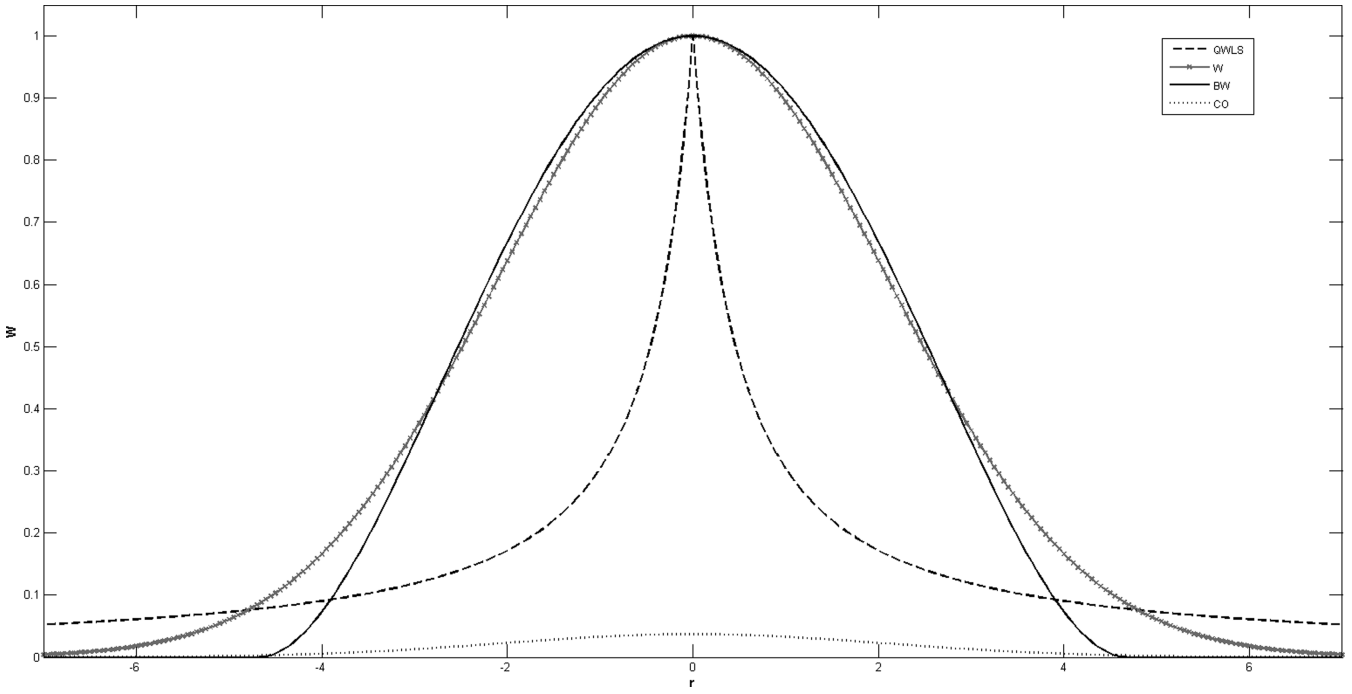

Figure 3. Weight functions.

when a monotone M-estimator is used. Furthermore, the solution of eq 8 is easier than the corresponding to eq 3 for a redescending M-estimator because monotone estimators have a unique solution. Thus, the values used to start the iterative process may influence the number of iterations but not the final outcome. The sequential solution of eqs 6 and 8 do not coincide exactly with that of eq 3, but the difference is negligible for practical purposes.

The SoM incorporates an extra step to the SiM as follows.

2.1.3. Step 3. In this case, all the measurements in the time horizon are used for the estimation problem. Used as initial point the solution of eq 8, the estimator $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$, that satisfies eq 1, is the solution of

$$\sum_{j=t-N+1}^{t} \sum_{i=1}^{n} \rho_{\mathrm{BW}}\left(\frac{y_{ij} - x_i}{\sigma_i}\right) = \min \tag{10}$$

whose solution is the final estimate.

Step 3 is a refinement stage of the solution attained in Step 2. Given that the BW function is a bounded redescending estimator, which have different local optima, the initial point of the optimization problem must be robust to ensure the convergence to a good solution.

2.2. Welsch M-Estimator. The W M-estimator was introduced by Dennis and Welsch[21] as a soft redescending estimator, less sensitive to outliers than the monotone ones. Its loss function is represented by eq 11, and its influence function asymptotically approaches zero for large values of r, as can be seen from Figure 2:

$$\rho_{\mathrm{W}} = {c_{\mathrm{W}}}^2\left\{1 - \exp\left[-\left(\frac{r}{c_{\mathrm{W}}}\right)^2\right]\right\} \tag{11}$$

The function has been only applied to reconcile the measurements of dynamic processes[11,12] for time varying moving windows. Regarding the initialization of the estimation problems, measurement values are used as the initial guesses of the independent input variables, and the starting points for model parameters are problem dependent.

2.3. Quasi-Weighted Least Square M-Estimator. The loss function of this M-estimator[16] is defined as follows:

$$\rho_{\mathrm{QWLS}} = \frac{r^2}{2 + c_{\mathrm{QWLS}}|r|} \tag{12}$$

The addition of the term $c_{\mathrm{QWLS}}|r|$ to the denominator of the LS function reduces the effect of large errors, where c_{QWLS} is an adaptive tuning constant. This function is a monotone estimator, and $\psi_{\mathrm{QWLS}} \to 1/c_{\mathrm{QWLS}}$ when $r \to \infty$.

The QWLS estimator was used to address linear steady state data reconciliation problems, but no discussion was provided about the effect of the optimization problem initialization on the computation load.

2.4. Correntropy M-Estimator. The CO M-estimator[17] is defined by eq 13. The Gaussian kernel function depends on its kernel width c_{CO}. It was proposed to adjust the value of c_{CO} for the process under analysis to scale-up outliers. The influence function of CO tends quickly to zero for $|r| > c_{\mathrm{CO}}$:

$$\rho_{\mathrm{CO}} = \frac{1}{c_{\mathrm{CO}}\sqrt{2\Pi}} \exp\left[-\left(\frac{r^2}{2{c_{\mathrm{CO}}}^2}\right)\right] \tag{13}$$

Also, steady state data reconciliation problems were solved using this estimator as objective function. It was proposed to formulate the unconstrained optimization problem at first[2] and solve it using an iterative rewheiting procedure. The solution provided by the LS estimator was used as the starting point of the optimization problem.

3. PERFORMANCE ANALYSIS

In this work, the capabilities of the selected strategies to estimate variables and identify outliers are analyzed for two benchmarks.

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

The procedure performances are examined for three different measurement models:

(1) Model with no outliers. It is assumed that the standardized residuals have the same distribution F, represented by the standard normal distribution $F \sim N(0,1)$ in this case.
(2) Model with occasional outliers, represented by a heavy-tailed symmetric F. It is chosen as a contaminated normal distribution: $F \sim (1-\varepsilon)N(0,1) + \varepsilon N(0,K^2)$, where ε denotes the contamination rate. That is, with probability ε, a normal error is multiplied by a constant K. The parameter ε is set at 0.1 in this study.
(3) Model of failure. Most of the observations follow case 1, but a random proportion ε does not obey that model at all. Among the several possible scenarios of failure, the error is represented as a fixed value $R\sigma_i$.

In cases 1 and 2, the estimators have no bias, and therefore the MSE reflects only the variability. It is desirable to have estimators that have a high efficiency in both cases 1 and 2. Recall that the efficiency of a given estimator at a given distribution F is the ratio between the variances of the maximum likelihood estimator corresponding to F and of the given estimator. In case 3, the MSE reflects both the variance and a bias, which must be controlled.

To compare the estimation capabilities of different techniques, the efficiencies of the M-estimators are fixed at 95.5% at the ideal distribution by properly tuning their parameters. Ozyurt and Pike[10] dealt with this issue in the same way. For QWLS and CO estimators, the tuning is performed using the jackknife procedure.[22] Table 1 presents the parameter values for each loss function.

Table 1. Loss Functions—Tuning Parameters (Relative Efficiency = 0.95)

c_{CO}	c_{QWLS}	c_W	c_{BW}
2.05	0.89	2.98	4.68

Adaptive estimators intend to minimize the estimator's variance in cases 1 and 2. They use information from the sample to optimize the estimator's parameters, such as the constant c in eqs 12 and 13. There exist different ways to deal with this topic, that is, minimizing an estimate of the estimator's variance, maximizing the generalized T likelihood function evaluated at the initial estimates of the process states, using an iterative procedure based on the Akaike information criterion, etc. However, an adaptive estimator is not necessarily better than a properly tuned M-estimator. The following issues support this statement:[13]

(1) Extensive simulations in robust statistics[6] have shown that adaptive estimators, despite their greater computational complexity, do not outperform good estimators with cleverly chosen fixed parameters.
(2) The biweight estimator tuned to have efficiency 0.95 for normal F (which means taking $c_{BW} = 4.68$), for example, has an efficiency of 0.70 when F is the Cauchy distribution (i.e., distribution T with one degree of freedom), which is an extreme case of heavy-tailedeness.
(3) It must be recalled that the true value of the "optimal" parameter, such as c, is unknown for adaptive estimators. Only an estimate of it is available that has a certain bias and variance, which in turn are propagated to the reconciliation estimator. Therefore, this approach is reliable only with very large sample sizes. Recall that the MSE of an estimator can be decomposed as MSE = bias^2 + variance. Typically, the variance decreases as 1/(sample size), while the contamination bias in case 3 does not. Therefore, for large sample sizes where adaptive estimators make sense, efficiency is not as important as bias.

Ten-thousand simulation trials are performed for each case study, and the length of the data horizon is fixed at $H = 10$. The measures of performance used in this analysis are MSE, AVTI, and OP, which are estimated as follows:

$$\mathrm{MSE} = \frac{1}{nNs}\sum_{k=1}^{\mathrm{SiM}}\sum_{i=1}^{n}\left(\frac{\hat{x}_i - x_i}{\sigma_i}\right)^2 \tag{14}$$

$$\mathrm{AVTI} = \frac{\#(\text{gross errors incorrectly identified})}{Ns} \tag{15}$$

$$\mathrm{OP} = \frac{\#(\text{gross errors correctly identified})}{\#(\text{gross errors simulated})} \tag{16}$$

where Ns is the number of simulation trials.

Furthermore, the cutoff point of each technique, that is, the value beyond which the measurements are considered as outliers, is adjusted by trial and error in such a way that the AVTI is approximately 0.05 when no outliers are present, and measurements are generated using a normal distribution (case study 1). This practice comes from the earliest works in data reconciliation[23] and guarantees that all procedures have the same behavior when there are no outliers.

Regarding the starting point of the optimization problem, Chen et al.[17] reported that the CO estimator was initialized using the solution of the LS technique, but any reference was provided by Zhang et al.[16] with respect to this issue for the QWLS estimator. Given that CO and QWLS estimates have been recently compared for the SMN benchmark by Chen et al.,[17] the initialization used by the last authors is assumed for both procedures in this work. Furthermore, the same starting point is used for the W estimator because no mention appears in the literature about the initialization of the steady state estimation problem.

The procedures were executed using a Processor Intel Core (TM) i7 CPU 930 @ 2.80 GHz, 8GB RAM, using the Successive Quadratic Programming code of MatLab Release 7.12 (R2011a).

4. RESULTS

Next, simulation results are reported and analyzed in detail for each benchmark.

4.1. Steam Metering Network. The SMN, presented by Serth and Heenan,[15] involves 28 streams that interconnect 11 units. The flow rates of all streams are measured. Random errors are generated considering that the standard deviations of the observations are 2.5% of their true values. Tables 2 and 3 display the cutoff points of the methodologies and the performance measures for case study 1, respectively.

Table 2. Cutoff Points—SMN

SiM	SoM	CO	QWLS	W
3.84	3.8416	3.8165	3.753	3.811

Regarding case study 2, Table 4 includes the performance measures for different K values, but only the AVTI and MSE records are displayed in Figure 4 because the OPs are similar for all techniques. In Table 5, the averages of the execution times for 10 000 simulations are reported.

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 3. Results for Case Study 1—SMN

AVTI					MSE × 10^2				
SiM	SoM	CO	QWLS	W	SiM	SoM	CO	QWLS	W
0.0499	0.0499	0.05	0.0499	0.0499	6.384	6.387	6.4129	6.4109	6.3796

Table 4. Results for Case Study 2—SMN

	K	2	5	10	14	15	18	20
AVTI	SiM	0.052	0.053	0.049	0.051	0.048	0.049	0.046
	SoM	0.053	0.052	0.049	0.050	0.048	0.048	0.046
	CO	0.050	0.051	0.050	0.063	0.068	0.102	0.126
	QWLS	0.048	0.054	0.055	0.058	0.060	0.056	0.061
	W	0.050	0.051	0.050	0.060	0.065	0.097	0.117
OP	SiM	0.055	0.438	0.698	0.779	0.791	0.817	0.825
	SoM	0.055	0.438	0.698	0.779	0.791	0.817	0.825
	CO	0.055	0.438	0.698	0.778	0.791	0.816	0.824
	QWLS	0.055	0.437	0.697	0.778	0.791	0.816	0.824
	W	0.055	0.438	0.698	0.778	0.791	0.816	0.824
MSE × 10^2	SiM	7.531	7.921	7.643	7.518	7.473	7.445	7.411
	SoM	7.509	7.871	7.606	7.485	7.446	7.416	7.393
	CO	7.489	7.922	7.699	8.197	8.555	12.547	20.678
	QWLS	7.510	9.223	10.322	10.697	10.822	10.925	10.997
	W	7.458	7.917	7.684	8.097	8.426	12.425	17.808

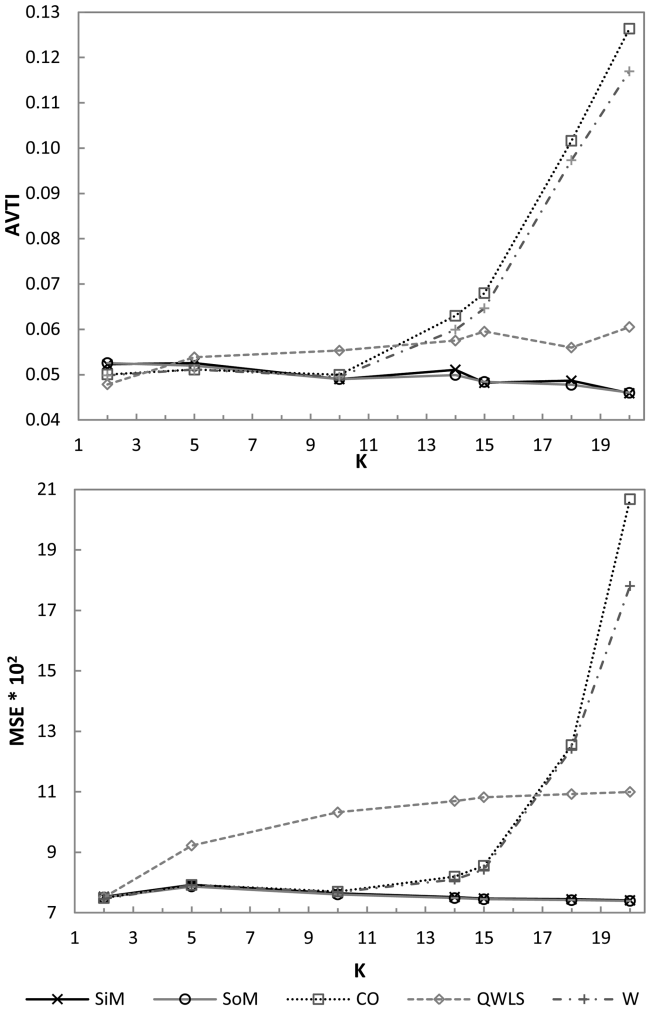

Figure 4. AVTI and MSE for case study 2—SMN.

Table 5. Average Execution Times (s) for Case Study 2—SMN

SiM	SoM	CO	QWLS	W
147.8	374.0	302.2	274.0	227.8

From Figure 4, it can be seen that

(a) The performances of the strategies SiM, SoM, W, and CO are similar for $K \in [2, 10]$, but SiM and SoM outperform CO and W for $K > 10$.
(b) The behaviors of CO and W are comparable for $K > 10$ even though there is evidence of a slight superiority of W with respect to CO.
(c) The AVTI and MSE values obtained for the QWLS increase with *K* and are poorer than those provided by SiM and SoM, except for $K = 2$. In contrast, QWLS behaves better than W and CO for large contaminations.
(d) The performance values of SiM and SoM are only slightly affected by *K*.

The analysis of Table 4 indicates that SoM provides a little better OP and MSE values with respect to SiM. The price for the improved performance of SoM is an increase in its computing time as a result of its extra step (Table 5).

The previous results point out that SiM works well for variable estimation and outlier detection for the whole range of *K* values and has the lowest computational requirements.

Next, the results for case study 3 are presented. Table 6 contains the performance measures for different *R* values. Also, the AVTI and MSE records are illustrated in Figure 5, and the average execution times are reported in Table 7.

If measurements do not obey the normal contaminated distribution, it can be seen from Figure 5 that

(a) The AVTI and MSE values for the QWLS increase with *K*.
(b) The performance indices for CO and W are comparable for all *R* values.

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 6. Results for Case Study 3—SMN

	R	1	2	3	4	5	6	7	8	9	10
AVTI	SiM	0.049	0.069	0.091	0.092	0.057	0.048	0.048	0.048	0.048	0.048
	SoM	0.049	0.069	0.088	0.083	0.056	0.048	0.048	0.048	0.048	0.048
	CO	0.049	0.067	0.080	0.073	0.056	0.048	0.047	0.047	0.047	0.047
	QWLS	0.047	0.061	0.072	0.082	0.090	0.097	0.101	0.106	0.109	0.111
	W	0.049	0.067	0.080	0.074	0.057	0.048	0.047	0.047	0.047	0.047
OP	SiM	0.000	0.000	0.002	0.584	0.998	1.000	1.000	1.000	1.000	1.000
	SoM	0.000	0.000	0.002	0.607	0.998	1.000	1.000	1.000	1.000	1.000
	Chen	0.000	0.000	0.001	0.541	0.996	1.000	1.000	1.000	1.000	1.000
	QWLS	0.000	0.000	0.000	0.384	0.971	0.999	1.000	1.000	1.000	1.000
	W	0.000	0.000	0.001	0.528	0.995	1.000	1.000	1.000	1.000	1.000
MSE × 10^2	SiM	7.681	12.539	15.092	11.601	7.545	7.167	7.166	7.166	7.166	7.166
	SoM	7.738	12.522	14.491	10.863	7.437	7.153	7.153	7.153	7.153	7.153
	CO	7.850	12.478	13.969	11.176	8.495	7.527	7.273	7.201	7.184	7.181
	QWLS	8.150	11.784	14.363	16.081	17.234	18.028	18.590	19.000	19.305	19.539
	W	7.782	12.406	14.160	11.545	8.677	7.565	7.261	7.172	7.149	7.144

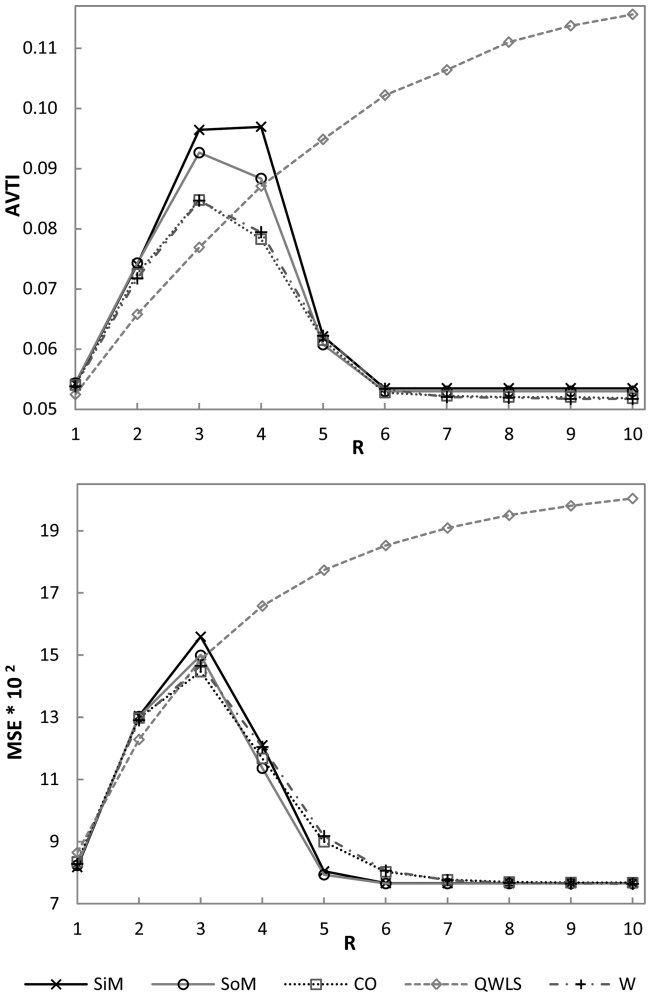

Figure 5. AVTI and MSE for case study 3—SMN.

Table 7. Average Execution Times for Case Study 3—SMN

SiM	SoM	CO	QWLS	W
132.3	356.8	314.2	274.3	231.0

(c) AVTI records for SiM and SoM are greater than the corresponding ones to W and CO for $R \in [1,...,4]$. For $R \geq 5$, SiM and SoM tend to the behavior of W and CO for the AVTIs.

(d) The MSE achieved using SoM is lower than the one attained by W and CO for $R = 1, 4-8$, and it is a little better than the MSE obtained using SiM.

From the analysis of the OP values reported in Table 6, it can be concluded that the performance of all techniques is similar, except for $R = 4$. In this case, the methodologies based on the BW function present higher OP values. Regarding the average computational times, Table 7 shows that the computational requirements of SoM are the highest, and those of SiM are the lowest.

Figure 5 shows that the AVTI and MSE of all redescending M estimators change with the increment of R in a similar way for case study 3. Taking into account the performance measures, there is no evidence of a clear superiority among the analyzed methodologies. It can be noticed that the SiM technique provides a good balance between the estimation and outlier detection/identification capabilities of the procedure and its computational time requirement.

The following comments arise from the analysis of the previous results:

(a) Even though the strategies based on QWLS, CO, and W functions have the same initialization (LS), the behavior of QWLS technique is different from the others two methods because QWLS function is a monotone M estimator.

(b) Both the W and CO functions are redescending M estimators formulated in terms of residual exponential functions. Therefore, they behave in a similar way.

(c) The BW function rejects outliers if the residuals are greater than 4.68; thus, the performance measures of SiM and SoM do not change significantly for $R > 5$.

Table 8. Cutoff Points—P&F

SiM	SoM	CO	QWLS	W
3.312	3.315	3.308	3.278	3.3028

4.2. Nonlinear Example. The second benchmark is extracted from the work by Pai and Fisher,[24] and it is symbolized as P&F. It involves six nonlinear equality constraints, which are defined in terms of five redundant measured and three observable unmeasured variables. Random errors are generated

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 9. Results for Case Study 1—P&F

AVTI					MSE × 10^2				
SiM	SoM	CO	QWLS	W	SiM	SoM	CO	QWLS	W
0.05	0.05	0.05	0.05	0.05	4.223	4.218	4.235	4.229	4.212

Table 10. Results for Case Study 2—P&F

	K	2	5	10	14	15	18	20
AVTI	SiM	0.036	0.035	0.034	0.039	0.034	0.034	0.036
	SoM	0.037	0.035	0.038	0.034	0.032	0.036	0.036
	CO	0.035	0.037	0.037	0.066	0.081	0.184	0.120
	QWLS	0.035	0.039	0.041	0.058	0.060	0.087	0.132
	W	0.037	0.035	0.035	0.043	0.051	0.081	0.127
OP	SiM	0.068	0.310	0.419	0.452	0.453	0.462	0.468
	SoM	0.071	0.311	0.418	0.448	0.453	0.468	0.470
	CO	0.071	0.307	0.415	0.452	0.454	0.469	0.472
	QWLS	0.071	0.306	0.416	0.445	0.456	0.462	0.477
	W	0.070	0.309	0.416	0.450	0.455	0.469	0.471
MSE × 10^2	SiM	4.837	5.010	4.676	4.679	4.674	4.577	4.588
	SoM	4.771	4.877	4.722	4.613	4.565	4.537	4.540
	CO	4.774	4.969	4.818	6.335	7.018	12.541	10.697
	QWLS	4.905	5.743	6.123	7.252	7.377	8.940	11.400
	W	4.827	4.886	4.810	4.943	5.525	7.0134	9.491

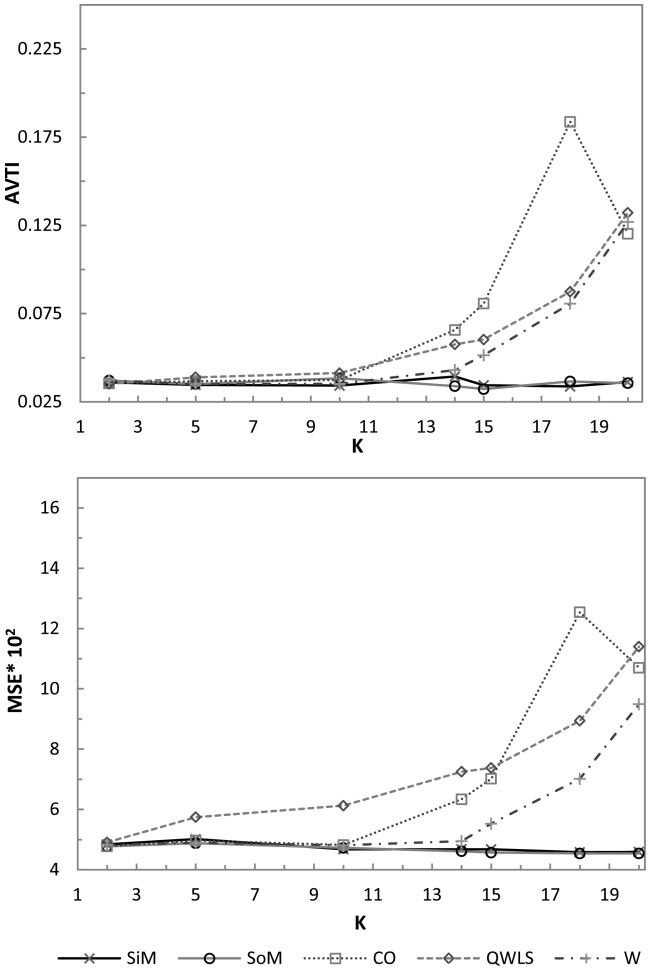

Figure 6. AVTI and MSE for case study 2—P&F.

considering the standard deviations suggested in that work. The same type of analysis provided for the linear example is presented for the nonlinear one.

Table 11. Average Execution Times (s) for Case Study 2—P&F

SiM	SoM	CO	QWLS	W
221.3	372.1	896.2	808.3	809.9

Tables 8 and 9 display the cutoff points of the methodologies and the performance measures for case study 1, respectively.

With respect to case study 2, Table 10 shows the performance measures for different *K* values, but only the AVTI and MSE records are displayed in Figure 6 because the OPs are similar for all techniques. Table 11 is composed of the average of execution times for 10 000 simulations.

From Figure 6, it can be observed that

(a) The performance measures of SiM and SoM are only slightly affected by *K*.
(b) SiM, SoM, CO, and W present the same behavior for $K \in [2,10]$, but the AVTI and MSE of CO and W increase for $K > 10$ and $K > 14$, respectively.
(c) In general, the QWLS function shows the poorest behavior regarding the MSE.
(d) The OP values are similar for all the analyzed techniques.

In contrast to the linear case, the use of the LS estimate as initialization of the robust estimation problem, as was suggested by Chen et al.,[17] increases the computational time in comparison to SOM's requirements (see Table 11).

Next, the results for case study 3 are presented. Table 12 contains the performance measures for different *R* values. Also, the AVTI and MSE records are illustrated in Figure 7, and the average execution times are reported in Table 13.

For case study 3, the results of the nonlinear example provide the same conclusions obtained for the linear benchmark. Taking into account the values of the performance measures, no clear superiority of one technique over another one can be verified for the studied range of *R* values. Regarding the computational time, the requirements of the SiM procedure are the lowest. Furthermore, the execution time of the strategies that use the

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 12. Results for Case Study 3—P&F

	R	1	2	3	4	5	6	7	8	9	10
AVTI	SiM	0.034	0.040	0.048	0.048	0.037	0.035	0.034	0.034	0.037	0.035
	SoM	0.035	0.040	0.047	0.043	0.038	0.036	0.035	0.036	0.036	0.034
	CO	0.034	0.039	0.042	0.047	0.044	0.040	0.035	0.038	0.036	0.046
	QWLS	0.036	0.036	0.039	0.049	0.052	0.060	0.067	0.064	0.068	0.071
	W	0.033	0.038	0.043	0.051	0.045	0.040	0.034	0.033	0.034	0.035
OP	SiM	0.000	0.000	0.074	0.953	0.999	1.000	1.000	1.000	1.000	1.000
	SoM	0.000	0.000	0.073	0.968	0.999	1.000	1.000	1.000	1.000	1.000
	CO	0.000	0.000	0.071	0.963	0.995	1.000	1.000	1.000	1.000	1.000
	QWLS	0.000	0.000	0.070	0.949	0.988	0.999	1.000	1.000	1.000	1.000
	W	0.000	0.000	0.073	0.964	0.995	1.000	1.000	1.000	1.000	1.000
MSE × 10^2	SiM	4.392	6.638	9.188	8.395	5.087	4.625	4.665	4.575	4.646	4.671
	SoM	4.463	6.572	8.636	7.394	4.828	4.656	4.527	4.562	4.633	4.622
	CO	4.521	6.713	7.923	8.020	6.274	5.181	4.693	4.808	4.789	5.191
	QWLS	4.817	6.219	7.565	9.033	10.036	10.683	10.942	11.492	11.455	11.897
	W	4.522	6.463	7.865	7.895	6.312	5.186	4.656	4.567	4.563	4.593

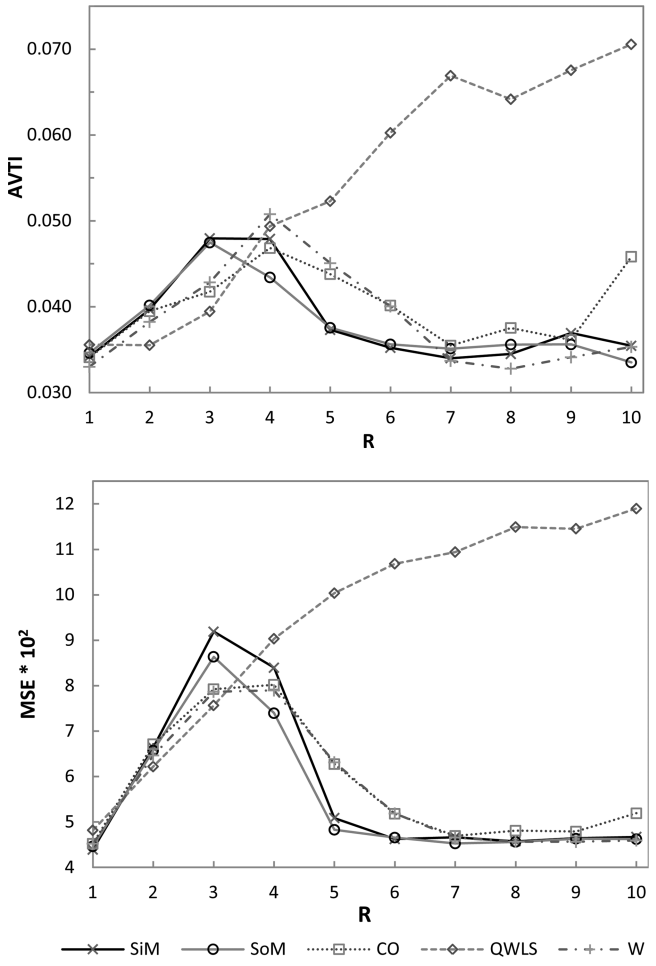

Figure 7. AVTI and MSE for case study 3—P&F.

Table 13. Average Execution Times (s) for Case Study 3—P&F

SiM	SoM	CO	QWLS	W
230.6	384.3	770.1	621.3	615.1

LS estimate as initial point is greater than the time consumed by SoM.

5. CONCLUSIONS

In this work, a performance comparison of five strategies for solving the robust data reconciliation problem is presented. Those techniques make use of the M estimators that have appeared in the data reconciliation literature during the past decade. All strategies are tuned to have the same performance when outliers are not present. The behaviors of the methodologies are analyzed for two measurement error models.

Results show that monotone and redescending M estimators behave differently even though the same initialization of the optimization problem is used. In this sense, the AVTI and MSE for QWLS increment in general for increasing values of contamination, in contrast CO and W, are more robust.

When measurement errors come from a contaminated normal, SiM and SoM are more robust than W and CO M-estimators for all the tested values of contamination. If those errors do not obey the aforementioned distribution, the performance measures of redescending M estimators change with contamination in a like manner, and no clear superiority of one technique over the other ones can be established.

In general, the lowest MSE is achieved using SoM, and SiM consumes the lowest computational time. Taking into account the trade-off between performance measures and computational work, SiM procedure appears as an efficient alternative for solving the type of problems under analysis. It provides good estimates for the reconciled measurements, and its computational load is the lowest thanks to the benefits of the robust initialization of the reconciliation problem performed running the Step 1 of the procedure.

The simultaneous treatment of outliers and gross errors, like biases, requires a different strategy, and it will be the subject of future works.

AUTHOR INFORMATION

Corresponding Author

*E-mail: msanchez@plapiqui.edu.ar.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors wish to thank the financial support of CONICET (National Research Council of Argentina), ANPCyT (National Agency of Scientific and Technological Promotion of Argentina),

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

UNS (Universidad Nacional del Sur, Argentina), and UBA (Universidad Nacional de Buenos Aires, Argentina).

REFERENCES

(1) Narasimhan, S.; Jordache, C. *Data Reconciliation and Gross Error Detection*; Gulf Publishing Company: Houston, TX, 2000.
(2) Romagnoli, J.; Sánchez, M. *Data Processing and Reconciliation for Chemical Process Operations*; Academic Press: San Diego, CA, 2000.
(3) Tjoa, I. B.; Biegler, L. T. Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems. *Comput. Chem. Eng.* **1991**, *15*, 679−690.
(4) Albuquerque, J. S.; Biegler, L. T. Data Reconciliation and Gross Error Detection for Dynamic Systems. *AIChE J.* **1996**, *42*, 2841−2856.
(5) Huber, P. J. *Robust Statistics*; Wiley: New York, 1981.
(6) Maronna, R.; Martin, R. D.; Yohai, V. *Robust Statistics: Theory and Methods*; John Wiley and Sons Ltd.: Chichester, England, 2006.
(7) Arora, N.; Biegler, L. T. Redescending Estimators for Data Reconciliation and Parameter Estimation. *Comput. Chem. Eng.* **2001**, *25*, 1585−1599.
(8) Hampel, F. R. The Influence Curve and Its Role in Robust Estimation. *J. Am. Stat. Assoc.* **1974**, *69*, 383−393.
(9) Wang, D.; Romagnoli, J. A. A Framework for Robust Data Reconciliation Based on a Generalized Objective Function. *Ind. Eng. Chem. Res.* **2003**, *42*, 3075−3084.
(10) Özyurt, D. B.; Pike, R. W. Theory and Practice of Simultaneous Data Reconciliation and Gross Error Detection for Chemical Processes. *Comput. Chem. Eng.* **2004**, *28*, 381−402.
(11) Martinez Prata, D.; Pinto, J. C.; Lima, E. L. Comparative Analysis of Robust Estimators on Nonlinear Dynamic Data Reconciliation. *Comput.-Aided Chem. Eng.* **2008**, *25*, 501−506.
(12) Martinez Prata, D.; Schwaab, M.; Lima, E. L.; Pinto, J. C. Simultaneous Robust Data Reconciliation and Gross Error Detection through Particle Swarm Optimization for an Industrial Polypropylene Reactor. *Chem. Eng. Sci.* **2010**, *65*, 4943−4954.
(13) Sánchez, M.; Maronna, R. Simple Approaches for Robust Data Reconciliation. *AIChE Annu. Meet., Conf. Proc.* **2009**, 79788.
(14) Maronna, R. A.; Arcas, J. Data Reconciliation and Gross Error Detection Based on Regression. *Comput. Chem. Eng.* **2009**, *33*, 65−71.
(15) Serth, R.; Heenan, W. Gross Error Detection and Data Reconciliation in Steam-Metering Systems. *AIChE J.* **1986**, *32*, 733−741.
(16) Zhang, Z.; Shao, Z.; Chen, X.; Wang, K.; Qian, J. Quasi-Weighted Least Squares Estimator for Data Reconciliation. *Comput. Chem. Eng.* **2010**, *34*, 154−162.
(17) Chen, J.; Peng, Y.; Munoz, J. Correntropy Estimator for Data Reconciliation. *Chem. Eng. Sci.* **2013**, *104*, 10019−10027.
(18) Nicholson, B.; López-Negrete, R.; Biegler, L. T. On-Line State Estimation of Nonlinear Dynamic Systems with Gross Errors. *Comput. Chem. Eng.* **2014**, *70*, 149−159.
(19) Zhang, Z.; Chen, J. Correntropy Based Data Reconciliation and Gross Error Detection and Identification for Nonlinear Dynamic Processes. *Comput. Chem. Eng.* **2015**, *75*, 120−134.
(20) Narasimhan, S.; Mah, R. S. H. Generalized Likelihood Ratio Method for Gross Error Identification. *AIChE J.* **1987**, *33*, 1514−1521.
(21) Dennis, J. E.; Welsch, R. E. Techniques for Nonlinear Least Squares and Robust Regression. *Proc. Am. Stat. Assoc.* **1976**, 83−87.
(22) Rey, W. J. J. *Introduction to Robust and Quasi-Robust Statistical Methods*; Springer-Verlag Berlin Heidelberg: Berlin, Germany,1983.
(23) Iordache, C.; Mah, R.; Tamhane, A. Performance Studies of the Measurement Test for Detection of Gross Errors in Process Data. *AIChE J.* **1985**, *31*, 1187−1201.
(24) Pai, D.; Fisher, G. Application of Broyden's Method to Reconciliation of Nonlinearly Constrained Data. *AIChE J.* **1988**, *34*, 873−876.

DOI: 10.1021/ie504735a
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX