Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 28 (2004) 381–402
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes

Derya B. Özyurt[1], Ralph W. Pike

to be random variables from the same distribution (univariate) with zero mean and unit deviation. Then similar to (6), but with a general matrix A for the linear case, additional constraints, and l_i 1, data reconciliation problem can be stated as:

$$\begin{aligned} &\min \sum_{i\ 1}^{n} \frac{(y_{i,1} - x_{i,1})^2}{\sigma_{i,1}^2} \quad \text{such that} \\ &Ax \quad 0, \\ &A \text{ is the process matrix} \\ &Lb \quad x \quad Ub \end{aligned} \tag{7}$$

Formulation (7) can be further generalized to include the unmeasured variables (u) and nonlinear process model constraints (f, g), which is frequently used in the data reconciliation literature.

$$\begin{aligned} &\min (y - x) Q^{-1} (y - x) \quad \text{such that} \\ &g(x, u) \quad 0 \\ &f(x, u) \quad 0 \\ &Lb_x \quad x \quad Ub \end{aligned} \tag{8}$$

where p_i is the probability and $b_i^2\sigma_i^2$ the variance of the contamination by a gross error.

For Logistic distribution, function (10) becomes

$$\max_i P_i \quad \max_i \frac{1}{\sigma_i} \frac{\exp((y_i - x_i)/\sigma_i)}{(1 \exp((y_i - x_i)/\sigma_i))^2} \quad \text{or}$$

$$\min_i \quad 2\ln 1 \exp \frac{(y_i - x_i)}{\sigma_i} - \frac{(y_i - x_i)}{\sigma} \ln \sigma_i \qquad (13)$$

Table 1
Tuning constants for different ρ functions with efficiency values 95.5%

ρ function	Tuning constants
Contaminated Normal	b_{CN} 10, p_{CN} 0.235
Cauchy	c_C 2.3849
Logistic	c_{Lo} 0.602
"Lorentzian"	c_L 2.6
Fair	c_F 1.3998
Hampel	a_H 1.35, b_H 2.7, c_H 5.4

To compare the data reconciliation and gross error detection performance of these ρ functions, they were first standardized by properly tuning their parameters. Some functions have their tuning constants given as a function of asymptotic efficiency such as the Fair and Cauchy functions. However, these asymptotic variances "give only crude indications for the actual variances" for finite sample size (Hampel, 2002). Therefore, approximate finite sample variances and consecutively relative efficiencies were calculated by simulation and Monte Carlo studies (Hampel, 1985; Andrews et al., 1972). We performed a similar study for the above ρ functions with a sample size of 28 and 2000 simulation runs that resulted in the following tuning constant values (efficiency values are approximately 95.5%) given in Table 1.

Fig. 1 depicts individual standardized ρ functions in the objective function, showing that Fair and Logistic functions cases result in a convex objective function. The convexity of the objective function guarantees the global optimality of the nonlinear data reconciliation problem for a process, which can be described by only linear constraints.

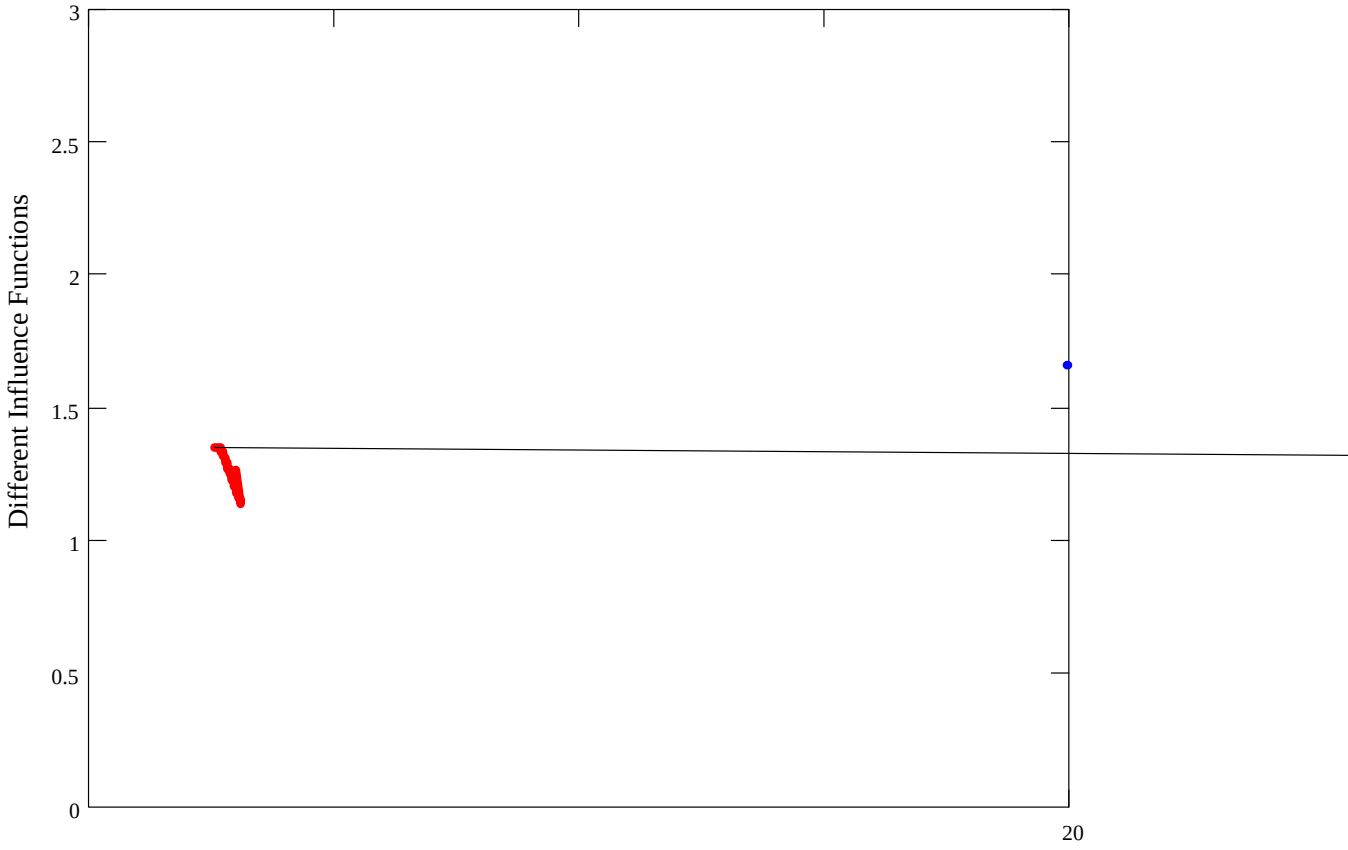
Methods to measure the robustness of an estimator involve the use of the influence function, IF (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986), which is defined for a sample x, an estimator T over an assumed distribution function F and a perturbed distribution function F_t as follows:

$$\mathrm{IF}(x, T, F) \quad \lim_{t \to 0} \frac{T(F_t) - T(F)}{t} \quad \frac{\partial}{\partial t}[T(F_t)]_{t=0} \qquad (23)$$

The heuristic interpretation of this influence function is that "it describes the effect of an infinitesimal contamination at the point x on the estimate" (Hampel et al., 1986). Since the influence function is proportional to the derivative of the maximum likelihood function, the weight given to any gross error in the measurements while calculating the estimates can be seen in Fig. 2 (see Appendix A for details).

The influence function for WLS is proportional to the measurement error (derivative of Eq. (12)) justifying the low breakdown point and unbounded effect of large errors. The effect of larger errors is reduced for the ρ function of the Cauchy distribution, "Lorentzian" function and Hampel's redescending M-estimator, shown by gradually decreasing influence functions in the region of greater than 3.0 of the standard error. Therefore, these three ρ functions are called

redescending efWLS w 9rofMC LS reduced0c >>2pel]TJ T [(ao)-21.1(i 5716 2.1(i 5Lom5716 2Ta6.rofMC 57bua)-93.WLS)-194lim e of*


Different Influence Functions
3
2.5
2
1.5
1
0.5
0
20

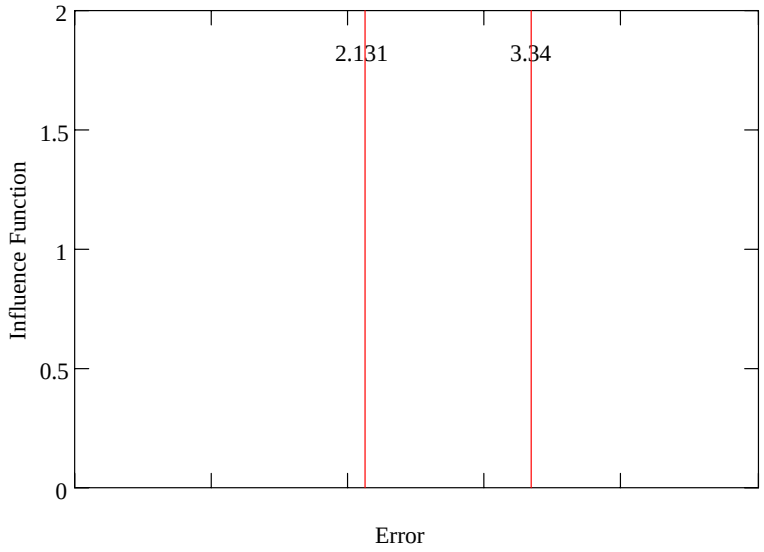


Fig. 3. Influence function for the ρ function of contaminated Normal distribution and five different cut points for gross error detection (first marker: maximum of the influence function (2.131), (□): inflection point of the first derivative of the influence function (2.42), (◇) Farris–Law criteria (2.65), (○) inflection point of the influence function (2.92); second marker: another inflection point of the first derivative of the influence function (3.34)).

CONOPT2 and MINOS5. In the first five examples, the piece-wise linear Hampel's redescending M-estimator is modeled as an external function coded in the programming language C and called by GAMS (Kalvelagen, 2002). For the last two problems, these discontinuities are smoothed as described in Arora and Biegler (2001). All calculations for the performance measures and the gross error detection rule X84 are implemented with Perl.

5.1. Examples from literature

The methods presented above are tested first on examples used in various literature articles in the last three decades. Two of these examples (Examples 1 and 2) contain linear and the remaining three (Examples 3-5) nonlinear process models. Except in Example 5, the lower bounds on the variables are set to 50% of the true values and the upper bounds to twice the true values. In Example 5, the lower bounds for all variables are 50% of the true values whereas the upper bounds are set to 150% of the true values.

Example 1 (Ripps, 1965). This example involves a simple chemical reactor with two entering and two leaving mass flows. All four variables are measured in the system, and they are related by three linear mass balance equations (Ripps, 1965; Romagnoli & Sanchez, 2000). For the Monte Carlo study, random measurements are created from Normal and Cauchy distributions as outlined above. Outliers were created in 10% of the measurements randomly by adding or subtracting 10–100% of the true values. With the exceptions of the Hampel's redescending M-estimator and MIMT, all runs were executed independently and with the same initial conditions. For MIMT, all consecutive runs were initiated with the resulting values of the previous run. Hampel's redescending M-estimator converged to an inferior optimal if it was not initialized with the results from Cauchy distribution ρ function or Fair function method.

The results of Monte Carlo study runs for each method are shown in Table 4. The ρ function of the Cauchy distribution shows the best performance with second highest overall power and lowest average number of Type I errors if the first cut point at 2.385 is used. Rule X84 seems to be conservative for this example, and the factor 5.2 can be reduced to improve the results. The comparison of the data reconciliation performance shows that ρ thic(res)238]TJ7

Table 4
Performance of different methods for Example 1

Table 5
Performance of different methods for Example 2

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	929	1100	1052	1000	1153	1000	1085
Total GE	6955	6456	7579	7415	6914	8053	6908	6824
Runs with GE	1000	929	1100	1052	999	1153	999	1084
OP (GED #1)	0.684	0.705	0.759	0.724	0.720	0.744	0.744	0.712
AVTI (GED #1)	1.364	2.118	7.645	3.371	2.255	4.692	4.193	3.253
OP (GED #2)	–	0.684	0.751	0.705	0.678	0.718	0.704	0.678
AVTI (GED #2)	–	1.826	7.296	3.203	1.500	4.144	2.622	2.038
OP (GED #3)	–	0.700	0.338	0.689	0.702	0.707	0.650	0.670
AVTI (GED #3)	–	2.713	0.882	3.281	2.421	4.846	2.499	2.699
Mean TER	0.558	0.505	0.412	0.455	0.525	0.384	0.494	0.460
Median TER	0.552	0.504	0.385	0.466	0.516	0.400	0.472	0.447

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number (*i* 1, 2, 3 for first and second cut points and rule X84, respectively).

Similar to Example 2, modified MIMT outperformed other methods in data reconciliation. Once again, the ρ function of Cauchy distribution shows that comparable, if not superior results can be achieved in a single NLP solution (see Table 6).

Example 4 (Pai and Fisher, 1988). In this example, there are six nonlinear equality constraints, five measured variables—all measurements are redundant—, and three observable unmeasured variables. On the average, 25% of the generated measurements are contaminated with gross errors ranging from 10 to 100% of the exact values reported in Pai and Fisher (1988).

As seen in Table 7, the ρ function of Cauchy distribution results in the highest total error reduction whereas the ρ function for contaminated Normal reaches the highest overall power but with more occurrences of Type I errors.

Example 5 (Swartz, 1989). Another widely used literature example is the nonlinear heat exchanger network problem described by Swartz (1989), and Romagnoli and Sanchez (2000). The system of four heat exchangers is modeled with 17 material and energy balances. The total number of variables in the system is 30, of which 16 are measured and the rest is unmeasured. There are 10 redundant and 6 non-redundant measured variables.

Gross errors are generated only for the redundant measured variables and on the average of 25% of the time. The magnitude of the errors range between 5 and 10 standard deviations for the flow rates and between 5 and 30 standard deviations for the temperature variables.

Most of the methods studied show poor data reconciliation results with close to none average total error reductions (Table 8). The ρ function for contaminated Normal and the "Lorentzian" function prove to be the best options for this case.

5.2. Industrial examples

Not many industrial examples have been investigated for the performance of different data reconciliation and gross error detection methods. The few cases in the open literature study industrial process subsystems such as reactors (Sanchez et al., 1996; Weiss et al., 1996), or utilize simulated plant measurements (Jordache et al., 2001) instead of real time plant data. The first industrial example involving

Table 6
Performance of different methods for Example 3

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	1077	1076	1028	1006	1110	1000	1026
Total GE	5986	6398	6389	6172	5990	6569	5935	5546
Runs with GE	1000	1076	1075	1027	1005	1109	999	1025
OP (GED #1)	0.744	0.776	0.843	0.758	0.774	0.757	0.822	0.799
AVTI (GED #1)	1.744	2.234	8.571	4.488	2.583	3.722	4.986	3.675
OP (GED #2)	–	0.760	0.836	0.742	0.733	0.722	0.779	0.764
AVTI (GED #2)	–	1.964	8.172	4.330	1.809	3.149	3.102	2.362
OP (GED #3)	–	0.736	0.209	0.607	0.703	0.642	0.585	0.667
AVTI (GED #3)	–	1.945	0.391	2.771	1.757	2.478	1.143	1.492
Mean TER	0.622	0.585	0.423	0.450	0.587	0.475	0.552	0.539
Median TER	0.625	0.579	0.400	0.477	0.583	0.511	0.538	0.526

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number (*i* 1, 2, 3 for first and second cut points and rule X84, respectively).

Table 7
Performance of different methods for Example 4

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	1058	1000	1032	1000	1000	1000	1018
Total GE	1265	1350	1265	1320	1265	1265	1265	1279
Runs with GE	771	824	771	804	771	771	771	785
OP (GED #1)	0.580	0.601	0.597	0.666	0.614	0.639	0.627	0.611
AVTI (GED #1)	0.225	0.341	0.322	0.411	0.280	0.342	0.351	0.330
OP (GED #2)	–	0.504	0.578	0.588	0.469	0.526	0.515	0.494
AVTI (GED #2)	–	0.278	0.271	0.315	0.136	0.186	0.159	0.161
OP (GED #3)	–	0.442	0.180	0.468	0.386	0.420	0.333	0.335
AVTI (GED #3)	–	0.298	0.160	0.280	0.235	0.250	0.232	0.247
Mean TER	0.538	0.321	0.493	0.369	0.542	0.478	0.526	0.511
Median TER	0.568	0.514	0.538	0.572	0.593	0.586	0.569	0.558

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number (*i* 1, 2, 3 for first and second cut points and rule X84, respectively).

real plant data and process model was given in Chen et al. (1998).

In this subsection, the sulfuric acid process from (Chen et al., 1998) and 1998

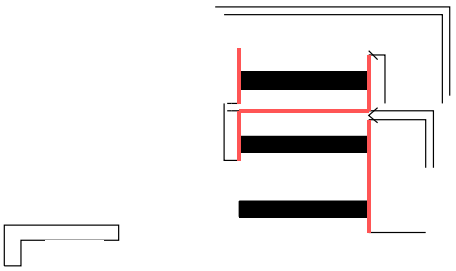


Table 9
Performance of different methods for Example 6

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	500	500	504	524	514	510	500	509
Total GE	3120	3392	3305	3346	3215	3181	3181	3026
Runs with GE	500	500	504	524	514	510	500	509
OP (GED #1)	0.841	0.863	0.884	0.912	0.889	0.904	0.907	0.896
AVTI (GED #1)	3.064	4.556	6.706	5.179	4.907	5.484	7.802	6.796
OP (GED #2)	–	0.833	0.879	0.897	0.848	0.876	0.866	0.844
AVTI (GED #2)	–	3.168	6.163	3.532	2.580	3.122	4.057	3.220
OP (GED #3)	–	0.819	0.670	0.871	0.827	0.852	0.778	0.784
AVTI (GED #3)	–	2.892	1.730	2.882	2.397	2.698	2.406	2.083
Mean TER	0.721	0.662	0.636	0.708	0.759	0.679	0.665	0.653
Median TER	0.767	0.714	0.661	0.778	0.802	0.779	0.682	0.689

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number (*i* 1, 2, 3 for fi

The alkylate product is a mixture of gasoline boiling range branched hydrocarbons which is blended with the refinery gasoline pool to increase the gasoline octane.

A simplified process flow diagram for a generic sulfuric acid alkylation process is given in Fig. 7. Specifically, Motiva alkylation process consists of five distinct sections, namely reaction, refrigeration, depropanizer, deisobutanizer and saturate deisobutanizer sections. The process has four reactor pairs and four acid settlers. In the reaction section, there are three feed streams: the olefin feed, the isobutane feed and the recycled olefin/isobutane mixture. The olefin feed contains the light olefins that are reacted with isobutane in the alkylation unit's STRATCO stirred reactors. The isobutane stream is in excess to fully react with all of the olefins being charged to the unit.

The alkylation process model developed using process flow diagrams, process data and process systems expertise has 1579 mostly nonlinear equality and 50 inequality constraints. The process model has 112–122 measured variables (122 for the first and second steady states, and 112 for the third steady state investigated in this study), 1512–1522 unmeasured variables and 67 parameters. The process measurements obtained from the distributed control system include 31 temperature, 30 flowrate, four pressure and 47–57 composition measurements. These measured variables, their standard deviations and the details of the model are given in Özyurt, Pike, Hopper, Punuru, and Yaws (2001), and Rich et al. (2001).

For the alkylation plant, three different steady-state operation points were determined from the data obtained on

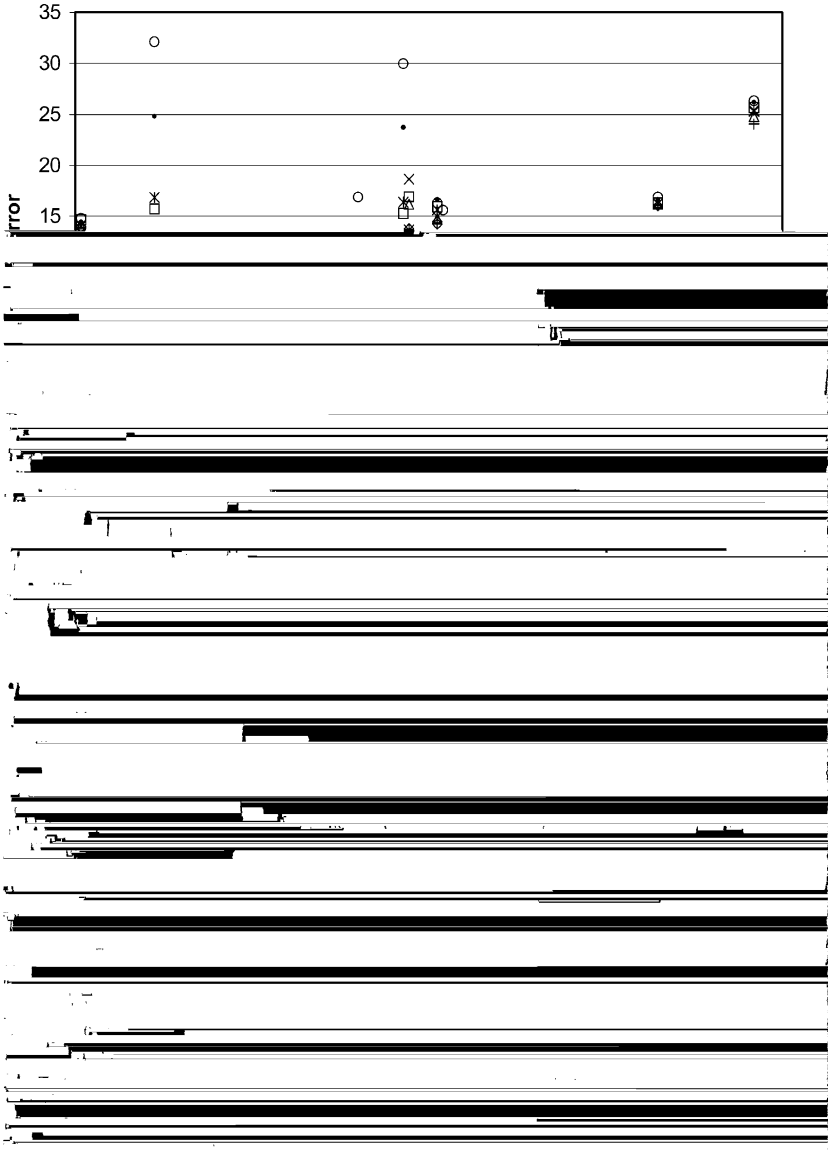


Fig. 9. Standard errors in measurements after reconciliation of the alkylation plant data at the second steady state: (a) all errors; (b) errors between –5 and 5. (◇) MIMT; ( ) Hampel; ( ) WLS; (□) CN; (✕) Cauchy; (○) Lorentzian; (×) Fair; ( ) Logistic.

8–9 November and 6–7 December 1998. The gross errors detected by the methods requiring single NLP solution vary between 19 and 44 for the first, between 25 and 43 for the second and between 23 and 41 for the third steady state (Table 11). MIMT-GED #1 and GED #2 for all other methods suggest that the second steady state has the most gross errors followed by the third steady-state operation point. Considering MIMT-GED #1, H-GED #2 and Cauchy-GED #2, the range for detected gross errors is 23–24, 28–30 and 26–27 for the first, second and third steady states, respec-

Terral values for

s for detected

able 1-113.8(measumentnd)]TJ T*5(rans.B.)Tj EMC Pan <</MCI9 1 >>BDC182 -1.2 Td St7(aahird 4.9* (normalized04)

OP (GED #1)
0.9
0.7
0.5
0.3
OP (GED #2)
0.9
0.7
0.5
0.3
OP (GED #3)
0.8
0.6
0.4
0.2
0.0

AVTI (GED #1)
0
2
4
6
8
AVTI (GED #2)
8
6
AVTI (GED #3)
5
4

error detection criteria based on estimation ef

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

Brooke, A., Kendrick, D., & Meeraus, A. (1992). *Release 2.25: GAMS: A user's guide*. Danvers, MA: Boyd & Fraser Publishing Co.

Chen, X. (1998). The optimal implementation of on-line optimization for chemical and refinery processes, Ph.D. Dissertation. Baton Rouge, LA 70803: Louisiana State University.

Chen, X., Pike, R. W., Hertwig, T. A., & Hopper, J. R. (1998). Optimal implementation of on-line optimization. *Computers and Chemical Engineering, 22*(Suppl.), S435–S442.

Crowe, C. M. (1986). Reconciliation of process flow rates by matrix projection. Part II. Nonlinear case. *AIChE Journal, 32*(4), 616–623.

Crowe, C. M., Garcia Campos, Y. A., & Hyrmak, A. (1983). Reconciliation of process flow rates by matrix projection. Part I. Linear case. *AIChE Journal, 29*(6), 881–888.

Deutsch, R. (1965). *Estimation theory*. Englewood Cliffs, NJ: Prentice-Hall.

Fair, R. C. (1974). On the robust estimation of econometric models. *Annals of Economic and Social Measurement, 3*, 667–677.

Farris, R. H., & Law, V. J. (1979). An efficient computational technique for generalized application of maximum likelihood to improve correlation of experimental data. *Computers and Chemical Engineering, 3*, 95–104.