

Article

A Modified Expectation Maximization Approach for Process Data Rectification

Weiwei Jiang¹, Rongqiang Li¹, Deshun Cao¹, Chuankun Li¹ and Shaohui Tao^{2,*} 

¹ State Key Laboratory of Safety and Control for Chemicals, SINOPEC Qingdao Research Institute of Safety Engineering, Qingdao 266071, China; jiangww.qday@sinopec.com (W.J.); lirq.qday@sinopec.com (R.L.); caods.qday@sinopec.com (D.C.); lick.qday@sinopec.com (C.L.)

² College of Chemical Engineering, Qingdao University of Science and Technology, Qingdao 266042, China

* Correspondence: tau@qust.edu.cn

Abstract: Process measurements are contaminated by random and/or gross measuring errors, which degenerates performances of data-based strategies for enhancing process performances, such as online optimization and advanced control. Many approaches have been proposed to reduce the influence of measuring errors, among which expectation maximization (EM) is a novel and parameter-free one proposed recently. In this study, we studied the EM approach in detail and argued that the original EM approach is not feasible to rectify measurements contaminated by persistent biases, which is a pitfall of the original EM approach. So, we propose a modified EM approach here to circumvent this pitfall by fixing the standard deviation of random error mode. The modified EM approach was evaluated by several benchmark cases of process data rectification from literatures. The results show advantages of the proposed approach to the original EM in solving efficiency and performance of data rectification.

Keywords: data rectification; expectation maximization; bias detection



Citation: Jiang, W.; Li, R.; Cao, D.; Li, C.; Tao, S. A Modified Expectation Maximization Approach for Process Data Rectification. *Processes* **2021**, *9*, 270. <https://doi.org/10.3390/pr9020270>

Received: 26 November 2020

Accepted: 5 January 2021

Published: 30 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the advancement of smart manufacturing, process measurements play a more and more important role in modern chemical manufacturing plants [1–3]. The measurements are unavoidably contaminated by random errors and often by large-sized gross errors, too, which degenerate performances of process monitoring, control and optimization strategies based on measurements [1]. To recover the true values of process variables from the contaminated measurements, many approaches to data rectification, i.e., reducing the random and gross errors simultaneously from the measurements, have been proposed since 1960s [2].

Traditionally, there are three ways of process data rectification, namely, statistical test [4,5], robust estimator [6,7] and mixed integer programming [8,9].

The first way identifies gross errors with a statistical test by assuming random errors follow a normal distribution [10], then a procedure of data reconciliation, i.e., solving a constrained least squares problem whose objective is minimizing the difference between the measured values and reconciled values satisfying process models, is carried out to estimate the true values of the measurements not contaminated by gross errors, while the true values of the measurements contaminated by gross errors are treated as unknown parameters to be estimated. Although the algorithmic parameters, such as critical values of a statistical test, can be chosen with clear statistical meanings, only one gross error can be identified at a time because of the smearing effect of a large-sized gross error, so the approaches of a statistical test must identify gross errors one by one and elegant frameworks must be designed to promise the performance of data rectification [11].

The second way is based on robust estimators [6], which can simultaneously reduce the influences of random and gross errors by solving a constrained nonlinear least squares

problem once. Different from the approaches of the statistical test described above, it is assumed that measurements contaminated by random and/or gross errors can be described by a heavy tail statistical distribution, such as contaminated normal [12], Cauchy [6], redescending [13], quasi weighted least squares (QWLS) [14] and correntropy [15] etc., which can effectively reduce the smearing effect of gross errors. The advantages of robust estimators for data rectification are: (1) gross errors can be identified with data reconciliation simultaneously; (2) the parameters of the robust estimators can be determined via Monte Carlo methods with clear statistical meanings [16] or online line search methods based on the Akaike information criterion (AIC) [13]. Currently, the robust estimators may be the most popular approach for process data rectification.

The third way is based on mixed integer programming (MIP) techniques [8–10], whose objective is to minimize the number of identified gross errors and the difference between the measured and rectified values, where the trade-off can be realized by the AIC [10,13] or a predetermined weighting factor of the objective function [9,17]. The MIP techniques show competitive or comparable performances to robust estimators for process data rectification, and the MIP technique based on the AIC is free from setting algorithmic parameters to balance the fitness and complexity of the model; although a critical value of identifying gross errors still needs to be determined, this value can be easily obtained from daily operation experiences of instrumentation engineers [10,17].

Recently, a novel way of process data rectification based on statistical inference was proposed, such as the approaches of Bayesian inference [18]. Being a widely used method of statistical inference, expectation maximization (EM) [19–25] has also been applied to process data rectification [26,27]. The statistical inference approaches are based on the Bayes rule [18], which infers the unknown parameters by combining the information from collected data (measurements) and the prior probability distribution of the inferred parameters. Although current works assume prior distribution before process data rectification, some reasonable prior information on the random and gross errors of measurements, such as standard deviation of random errors and occurrence of gross errors for a specified sensor, can be collected and modeled from the experiences of plant operators and historical process data [18,27,28]. So, the authors believe that the statistical inference approach to process data rectification deserves to be studied.

The established EM approach [26] is an interesting statistical inference approach because it has no algorithmic parameter to be determined before data rectification, but just assumes that measurement errors follow a finite Gaussian mixture distribution. The large number of parameters to be estimated with the EM algorithm [29] lead to its low-efficiency solving procedure, and from experiences of the authors, the original EM approach cannot be applied to rectify process measurements contaminated by persistent biases, because the estimated standard deviation of the random error mode is close to that of the gross error mode, which leads to difficulty of bias identification.

In this work, we argue that, for the original EM approach, the estimated value of standard deviation of random error mode is unavoidably enlarged by a persistent bias and leads to difficulty of bias detection. To circumvent this problem, we present a modified EM approach, where the standard deviation of random error mode is estimated before the EM iterations with a robust method [30], so the standard deviation of random error mode will not be enlarged by a persistent bias and it is possible to detect bias from the EM calculation result. Compared to the original EM approach, the modified one also reduces the number of parameters to be estimated and the time consumption of EM iterations can also be significantly reduced.

The remainder of this paper is organized as follows. Section 2 introduces the principles of the established EM approach for data rectification. The proposed modified EM approach and detailed calculation steps are presented in Section 3. Section 4 describes the performance analysis procedure used herein. The performance modified EM approach is evaluated and compared to the original one in Section 5. Finally, Section 6 concludes the paper.

2. Data Rectification Approach of Expectation Maximization

2.1. Data Rectification Problem

Except for the random errors following normal distribution, three types of gross errors, namely, drift, outlier and persistent bias, also usually contaminate measurements, which are shown in the following Figure 1.



Figure 1. Illustration of different types of gross errors.

Figure 1 shows how different types of gross errors contaminate a process measurement whose true value is 1 in steady state. Obviously, any one of the systematic errors significantly reduces the reliability of a process measurement, which is the basis of online decision-making during the enhancement of the process performance and a systematic error cannot be eliminated with data reconciliation methods, because a zero mean of random error is assumed for the methods of data reconciliation. Essentially, outliers in a measurement horizon are also random errors with larger variance than random noises, and the original EM algorithm can identify and estimate outliers well. A persistent bias as shown in Figure 1 is not random, which will enlarge the estimated variance of the original EM algorithm, as shown in the following Section 3.1. On drift error of a sensor, it is also a non-random one with increasing error size and it will also lead to an enlarged estimated variance as a persistent error, supposing an average of a measurement horizon is taken as a representative of the horizon. In the following, we show how a data rectification problem is set up as a statistical inference problem.

Supposing a measurement horizon $\{y_{j,h}\}_{h=t-H+1}^t$ is collected at time t , which involves H data points measured at different time point k for the j th process variable and a data matrix $Y \in R^{J \times H}$ whose rows represent measurement horizons of all the J measured process variables, where $y_{j,h}$ is an element at the j th row and h th column of Y . A steady-state process data rectification can be formulated as a maximum likelihood estimation problem described as the following Equation (1).

$$\begin{aligned} & \max_x \ln[P(\mathbf{Y}|\mathbf{x})] \\ & \text{s.t.} \begin{cases} f(\mathbf{x}) = 0 \\ g(\mathbf{x}) \leq 0 \end{cases} \end{aligned} \quad (1)$$

In Equation (1), the objective function is the logarithm likelihood of the sampled measurements $y_{j,h}$ under the condition that the true value of the j th measurement is x_j and the j th element of vector x is x_j ; $f(x)$ represents the process model and $g(x)$ denotes the

inequality constraints for the process variables, considering operational specifications and experienced bounds.

Assuming different distributions of the measurement errors, the formulation of the objective function of Equation (1) varies [6]. For the data reconciliation problem considering random errors only, it is assumed that random errors follow a normal distribution, and the logarithm of the objective function is a quadratic one. If a heavy tail distribution is assumed for measurement errors, as in the situation of data rectification using robust estimators, the logarithm of the objective function shows a more complex formulation, sometimes the function is nonconvex or even discontinuous [6].

In both the situations described above, we fix the distribution parameters of measurement errors. For the data rectification using the EM approach, the parameters of measurement error distribution are inferred with the Bayes rule, as described in the following section.

2.2. Expectation Maximization Approach

For the EM approach [26], the difference between the h th measurement and the true value of the j th process variable, namely, $\varepsilon_{j,h} = y_{j,h} - x_j$, is described with the finite Gaussian mixture model shown with Equation (2) [26].

$$\begin{aligned} \varepsilon_{j,h} &= y_{j,h} - x_j = w_{j,1}N(0, \sigma_{j,1}^2) + w_{j,2}N(0, \sigma_{j,2}^2) \\ \text{s.t.} \quad &\begin{cases} w_{j,1}, w_{j,2} \geq 0 \\ w_{j,1} + w_{j,2} = 1 \end{cases} \end{aligned} \quad (2)$$

In Equation (2), $w_{j,1}$ represents the probability of a random error mode with a zero mean and standard deviation $\sigma_{j,1}$, and $w_{j,2}$ represents the probability of a gross error mode with a zero mean and standard deviation $\sigma_{j,2}$ that is larger than $\sigma_{j,1}$ under the occurrence of a gross error. Supposing $\theta_j = \{\sigma_{j,k}, w_{j,k}\}_{k=1,2}$, the likelihood of $\varepsilon_{j,h}$ for the k th error mode can be described with the following Equation (3) [26].

$$p_{j,h,k} = P(y_{j,h} | z_{j,h} = k, x_j, \theta_j) = \frac{1}{\sqrt{2\pi\sigma_{j,k}^2}} \exp\left[-\frac{(y_{j,h} - x_j)^2}{\sigma_{j,k}^2}\right] \quad (3)$$

In Equation (3), $z_{j,h}$ is a latent variable to be estimated and $z_{j,h} = k$ represents that the error mode of $y_{j,h}$ is the k th one, where $k = 1$ represents a random error mode, and $k = 2$ means a gross error mode. Considering both error modes, the whole likelihood of $\varepsilon_{j,h}$ is represented as following Equation (4).

$$p_{j,h} = \sum_{k=1}^2 w_{j,k} p_{j,h,k} \quad (4)$$

Based on the above descriptions, the previously mentioned Equation (1) can be written as the following Equation (5).

$$\begin{aligned} \max_{x_j, \theta_j} \ln &\left[\prod_{j=1}^J \prod_{h=1}^H p_{j,h} \right] \\ \text{s.t.} \quad &\begin{cases} f(\mathbf{x}) = 0 \\ g(\mathbf{x}) \leq 0 \end{cases} \end{aligned} \quad (5)$$

It is difficult to solve Equation (5) directly, because $w_{j,k}$ cannot be obtained explicitly. Hence, an EM approach was applied to solve Equation (5), by replacing Equation (5) with the following Equation (6) [26].

$$\begin{aligned} \max_{\Theta} Q(\Theta, \Theta^{(t)}) &= E_{z_{j,h}=k|y_{j,h}, \Theta^{(t)}} \left[\ln \prod_{j=1}^J \prod_{h=1}^H p(z_{j,h} = k, y_{j,h} | \Theta) \right] \\ &= \sum_{j=1}^J \sum_{h=1}^H E_{z_{j,h}=k|y_{j,h}, \Theta^{(t)}} \ln p(z_{j,h} = k, y_{j,h} | \Theta) \\ &= \sum_{j=1}^J \sum_{h=1}^H \sum_{k=1}^2 p(z_{j,h} = k | y_{j,h}, \Theta^{(t)}) \cdot \ln p(z_{j,h} = k, y_{j,h} | \Theta) \\ \text{s.t. } &\begin{cases} f(\mathbf{x}) = 0 \\ g(\mathbf{x}) \leq 0 \end{cases} \end{aligned} \quad (6)$$

In Equation (6), $\Theta = \{x_j, \theta_j\}_{j=1, \dots, J}$ and $\Theta^{(t)}$ represent the estimation result of Θ at the t th iteration, which means the probability of error mode for a measurement, namely, $p(z_{j,h} = k | y_{j,h}, \Theta^{(t)})$, is estimated from $y_{j,h}$ and $\Theta^{(t)}$ using the Bayes rule, as described with Equation (7).

$$\begin{aligned} P(z_{j,h} = k | y_{j,h}, \Theta^{(t)}) &= \frac{P(z_{j,h}=k, y_{j,h}, \Theta^{(t)})}{P(y_{j,h}, \Theta^{(t)})} \\ &= \frac{P(y_{j,h}, z_{j,h}=k, \Theta^{(t)})}{P(y_{j,h}, \Theta^{(t)})} \\ &= \frac{P(y_{j,h} | z_{j,h}=k, \Theta^{(t)}) \cdot P(z_{j,h}=k, \Theta^{(t)})}{P(y_{j,h}, \Theta^{(t)})} \\ &= \frac{P(y_{j,h} | z_{j,h}=k, \Theta^{(t)}) \cdot P(z_{j,h}=k, \Theta^{(t)})}{P(y_{j,h}, \Theta^{(t)})} \\ &= \frac{P(y_{j,h} | z_{j,h}=k, \Theta^{(t)}) \cdot P(z_{j,h}=k | \Theta^{(t)}) \cdot P(\Theta^{(t)})}{P(y_{j,h} | \Theta^{(t)}) \cdot P(\Theta^{(t)})} \\ &= \frac{P(y_{j,h} | z_{j,h}=k, \Theta^{(t)}) \cdot P(z_{j,h}=k | \Theta^{(t)})}{P(y_{j,h} | \Theta^{(t)})} \\ &= \frac{p_{j,h,k} \cdot w_{j,k}}{\sum_{l=1}^2 p_{j,h,l} \cdot w_{j,l}} \end{aligned} \quad (7)$$

In Equation (7), $p_{j,h,k}$ is calculated with Equation (3) and $P(z_{j,h} = k | \Theta^{(t)}) = w_{j,k}$ because $\Theta = \{x_j, \theta_j\}_{j=1, \dots, J}$ and $\theta_j = \{\sigma_{j,k}, w_{j,k}\}_{k=1,2}$. The calculation of the probability of $P(z_{j,h} = k | y_{j,h}, \Theta^{(t)})$ is noted as the *expectation step* (E-step).

After the E-step, we estimate Θ using the *maximization step* (M-step), namely, solving Equation (6) with fixed $P(z_{j,h} = k | y_{j,h}, \Theta^{(t)})$ calculated at the E-step, then a new estimation of Θ , i.e., $\Theta^{(t+1)}$, is the result. It must be noted that $\ln p(z_{j,h} = k, y_{j,h} | \Theta)$ is calculated with Equation (3) and the Bayes rule, as described by Equation (8).

$$\begin{aligned} \ln p(y_{j,h}, z_{j,h} = k | \Theta) &= \ln \left[\frac{p(y_{j,h}, z_{j,h}, \Theta)}{p(\Theta)} \right] \\ &= \ln \left[\frac{p(y_{j,h} | z_{j,h}, \Theta) \cdot p(z_{j,h}, \Theta)}{p(\Theta)} \right] \\ &= \ln \left[p(y_{j,h} | z_{j,h}, \Theta) \cdot p(z_{j,h} | \Theta) \right] \\ &= \ln p_{j,h,k} + \ln \left[p(z_{j,h} | \Theta) \right] \\ &= -\ln \left(\sqrt{2\pi} \sigma_{j,k} \right) - \left(\frac{y_{j,h} - x_j}{\sigma_{j,k}} \right)^2 + \ln w_{j,k} \end{aligned} \quad (8)$$

To solve Equation (6), a coordinate search method is applied to estimate $w_{j,k}$ and $\sigma_{j,k}$ separately, as the following Equations (9) and (10) show [26].

$$\sigma_{j,k}^{(t+1)} = \sqrt{\frac{\sum_{h=1}^H \gamma_{j,h,k}^{(t)} \cdot (y_{j,h} - x_j^{(t)})^2}{\sum_{h=1}^H \gamma_{j,h,k}^{(t)}}}, \quad (9)$$

$$w_{j,k}^{(t+1)} = \frac{\sum_{j=1}^H \gamma_{j,h,k}^{(t)}}{H}. \quad (10)$$

In the above equations, $\gamma_{j,h,k}^{(t)} = P(z_{j,h} = k | y_{j,h}, \Theta^{(t)})$, which is calculated with Equation (7). At last, $x_j^{(t+1)}$ is estimated by solving Equation (6) with $w_{j,k}$ and $\sigma_{j,k}$ fixed as Equations (9) and (10) [26].

With the new estimation of Θ , i.e., $\Theta^{(t+1)}$, obtained, we return to the E-step and check the difference between $Q(\Theta, \Theta^{(t)})$ and $Q(\Theta, \Theta^{(t+1)})$ of Equation (6), if the difference is not obvious we stop the iteration, or the else we continue [26].

3. Modified Expectation Maximization Approach

3.1. Standard Deviation of the Original EM under Persistent Bias

Although the EM approach was successfully applied to several situations, such as non-persistent gross errors and concurrent errors of different types [26,27], there is still a little space for improvement in the situation of measurements contaminated by persistent biases, where $(y_{j,h} - x_j^{(t)})^2$ in Equation (9) is relatively large and unavoidably leads to a large $\sigma_{j,k}^{(t+1)}$ even for the random error mode, whose standard deviation shall be relatively small, as can be argued in the follows.

Supposing that $\alpha_{j,h,k} = \gamma_{j,h,k}^{(t)} / \sum_{h=1}^H \gamma_{j,h,k}^{(t)}$, Equation (9) can be rewritten as following Equation (11):

$$[\sigma_{j,k}^{(t+1)}]^2 = \sum_{h=1}^H \alpha_{j,h,k} \cdot (y_{j,h} - x_j^{(t)})^2, \alpha_{j,h,k} \in [0, 1] \text{ and } \sum_{h=1}^H \alpha_{j,h,k} = 1 \quad (11)$$

With $h' = \operatorname{argmin}_h \left\{ (y_{j,h} - x_j^{(t)})^2 \right\}$ and $\alpha_{j,h',k} = 1$, it is easy to infer that Equation (11) arrived at its minimum, namely, $[\sigma_{j,k}^{(t+1)}]_{\min} = |y_{j,h'} - x_j^{(t)}|_{\min}$, which fluctuates around the magnitude of the bias contaminating the j th measurement and leads to a large size standard deviation for random error mode of Equation (2), namely, $\sigma_{j,1}$. Under this situation, it is impossible to set $\pm 3\sigma_{j,1}$ as the critical value for bias detection and the original intention of Equation (2) is violated, too.

To verify the above argument, the simple linear data rectification case of Ripps [31] is used here to show the influence of persistent bias to the standard deviations of random mode. The Ripps case involves four streams with measured flowrates and three linear mass balance equality constraints are shown as the following Equation (12).

$$\begin{cases} 0.1x_1 + 0.6x_2 - 0.2x_3 - 0.7x_4 = 0 \\ 0.8x_1 + 0.1x_2 - 0.2x_3 - 0.1x_4 = 0 \\ 0.1x_1 + 0.3x_2 - 0.6x_3 - 0.2x_4 = 0 \end{cases} \quad (12)$$

The true values of all the flowrates are $x_1 = 0.1739$, $x_2 = 5.0435$, $x_3 = 1.2175$ and $x_4 = 4$, with corresponding standard deviations $\sigma_1 = 2.89 \times 10^{-4}$, $\sigma_2 = 2.5 \times 10^{-3}$, $\sigma_3 = 5.76 \times 10^{-4}$,

and $\sigma_4 = 4 \times 10^{-2}$ for random noises. Here we assume that x_3 is contaminated by a bias with sizes being $5\sigma_3$, $10\sigma_3$ and $15\sigma_3$, respectively, then 50 Monte Carlo simulations are carried for each bias size, with all the variables being added random noises with zero mean and corresponding standard deviations. For each Monte Carlo simulation, the sign of the bias is assigned randomly with equal probability. The minimum ratio of estimated $\sigma_{3,1}$ to σ_3 , namely, $\left(\frac{\sigma_{3,1}}{\sigma_3}\right)_{\min}$, is shown in Figure 2 as follows.

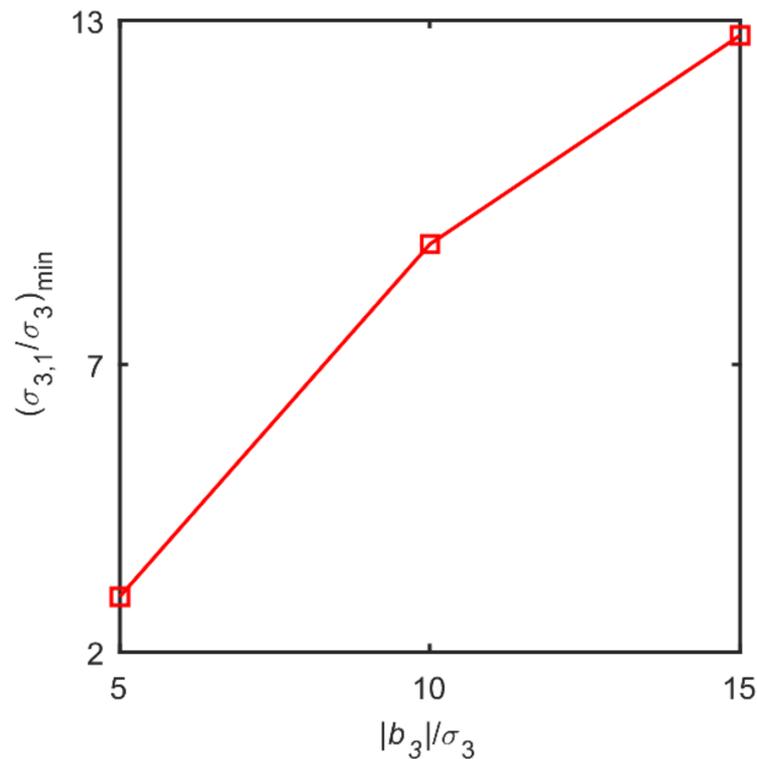


Figure 2. The influence of bias size on the estimated value of standard deviation of random error mode.

As Figure 2 shows, as in the above argument, the minimum estimated standard deviation of random noise mode is several times of that of the true, so it is impossible to detect a bias with the traditional 3σ rule, as the original EM approach did [26].

3.2. Modification to the Original EM Approach

To apply the EM approach to the situation of measurements contaminated by persistent biases, a simple modification is presented herein for the original EM approach, namely, we directly estimate the variance of the random error mode in Equation (2), i.e., $\sigma_{j,1}^2$, but not via the EM iterations. It has been shown that $\sigma_{j,1}^2$ can be estimated efficiently and robustly from process measurements even when the measurements are contaminated with gross errors [30]. Then all the other parameters of Θ in Equation (6) are still estimated using the above original EM procedure. Obviously, the influence of bias on the estimation of standard deviation of random error is avoided by this modification.

After Θ in Equation (6) being estimated, a criterion must be set up to detect bias for measurements. There are two established ways for detecting bias. The first one is shown by Equation (13), namely, a measurement is contaminated by a bias if the probability of gross error mode is larger than the random error mode [12]. The second one is using the deviation of reconciled value from the corresponding measured value [26], namely, if the

following Equation (14) holds where \bar{y}_j is the average of measurement horizon of the j th process variable, then a bias is detected for the j th measured variable.

$$w_{j,2} \sum_{h=1}^H \frac{1}{\sqrt{2\pi\sigma_{j,2}^2}} \exp \left[-\frac{(y_{j,h} - x_j)^2}{\sigma_{j,2}^2} \right] > w_{j,1} \sum_{h=1}^H \frac{1}{\sqrt{2\pi\sigma_{j,1}^2}} \exp \left[-\frac{(y_{j,h} - x_j)^2}{\sigma_{j,1}^2} \right], \quad (13)$$

$$|\bar{y}_j - x_j| > 3\sigma_{j,1}. \quad (14)$$

Obviously, the original EM approach can only use the first way of bias detection because the estimated value of $\sigma_{j,1}$ is enlarged by a persistent bias. While three criteria can be applied to the modified EM approach, i.e., a bias is detected when Equation (13) holds, which is noted as probability criterion (PC); or Equation (14) holds, which is noted as deviation criterion (DC); or both Equations (13) and (14) hold simultaneously, which is noted as a probability and deviation criterion (PDC).

Based on the above description, the proposed modified EM approach can be shown in Table 1.

Table 1. Modified expectation maximization (EM) approach for data rectification.

1. Input measurements matrix Y .
2. Estimate $\sigma_{j,1}^2$ from Y by using a robust direct approach [30].
3. Initialize parameters: $w_{j,1}^{(0)} = 0.5, \sigma_{j,2}^{(0)} = \sigma_{j,1}, x_j^{(0)} = \sum_{h=1}^H y_{j,h} / H$ and set $t = 1$.
4. **E-step.** Calculate $P(z_{j,h} = k | y_{j,h}, \Theta^{(t)})$ using Equation (7).
5. **M-step.** Calculate $\sigma_{j,2}^{(t+1)}$ and $w_{j,k}^{(t+1)}$ using Equations (9) and (10), respectively, calculate $x_j^{(t+1)}$ by solving Equation (6) with $\sigma_{j,2} = \sigma_{j,2}^{(t+1)}$ and $w_{j,k} = w_{j,k}^{(t+1)}$ fixed.
6. Terminate if $\left| \left[Q(\Theta, \Theta^{(t+1)}) - Q(\Theta, \Theta^{(t)}) \right] / Q(\Theta, \Theta^{(t)}) \right| \leq 10^{-3}$, or else $t = t + 1$ and return to step 4.
7. Detect bias for each measurement with PC, DC or PDC.

The modified EM with PC, DC or PDC for bias detection is noted as MEM-PC, EM-DC and EM-PDC, respectively.

The advantages of the modified EM algorithm over the original EM algorithm are: (1) the standard deviation of random error is not affected by a persistent bias, because a direct and robust variance estimation method [30] is used; (2) fewer variables need to be estimated by the modified EM algorithm, which means that the modified EM algorithm converges faster than the original EM algorithm.

4. Performance Analysis

To evaluate the performance of the proposed modified EM algorithm, the following three performance metrics, namely, overall performance (OP), average number of Type-I error (AVTI) and relative error reduction (RER), defined as following Equations (15)–(19) are used here [9].

$$OP = \frac{\text{number of correctly identified bias}}{\text{number of gross errors simulated}}, \quad (15)$$

$$AVTI = \frac{\text{number of wrongly identified bias}}{\text{number of simulation trials}}, \quad (16)$$

$$RER = \frac{\sum_{j=1}^M (MRE_j - RRE_j)}{\sum_{j=1}^M MRE_j}, \quad (17)$$

$$MRE_j = \left| \frac{\bar{x}_j - x_j^t}{x_j^t} \right|, \quad (18)$$

$$RRE_j = \left| \frac{x_j - x_j^t}{x_j^t} \right|. \quad (19)$$

The following Monte Carlo simulation procedure [6] is carried here to evaluate the performance of data rectification.

- (1) For all the measured variables, add random noises with zero mean and corresponding standard deviation.
- (2) Add bias to each measurement with a predefined probability p_b , the bias size randomly distributes in the range of 5 and 25 times of standard deviation of random noise, the sign of the bias, namely, '+' or '-', is randomly assigned with equal probability.
- (3) Calculate performance of data rectification with Equations (15)–(19) for each evaluated method.

Four well-known test cases of process data rectification were used here to evaluate and compare the performances of the proposed MEM approach to the original EM approach, which are described as following.

The first case is the famous steam metering network (SMN) [32], which involves 11 units interconnected by 28 streams with measured flow rates, whose flowsheet diagram is demonstrated as Figure 3 with the true values of the flowrates of all the streams shown in the parenthesis. For each measured variable, the standard deviation of added random noise is set as 2.5% of its true values.

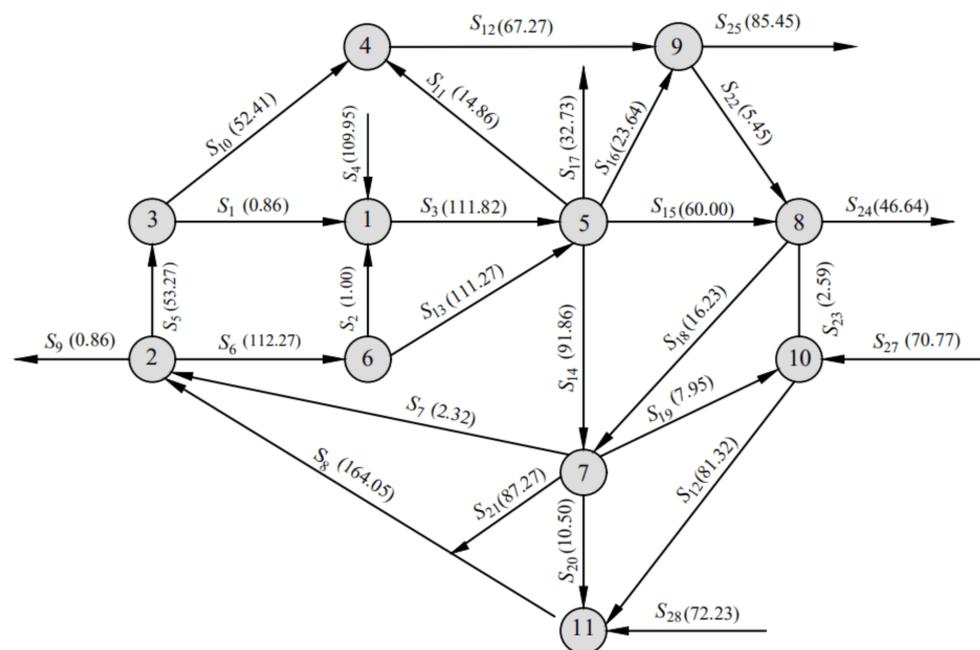


Figure 3. The steam metering network (SMN) flowsheet with true values of stream flowrates.

The second case is a bilinear process of metallurgical grinding (MG) [33], which involves four units interconnected by nine streams with measured mass flowrates and 15 measured mass fractions. The flowsheet of the metallurgical grinding is shown in Figure 4 with the true values of all the measured variables, where the true values of flowrates are shown in the parenthesis and composition shown at the right side of the parenthesis. For all the measured variables, the corresponding σ of random noise is set as 2.5% of its true value.

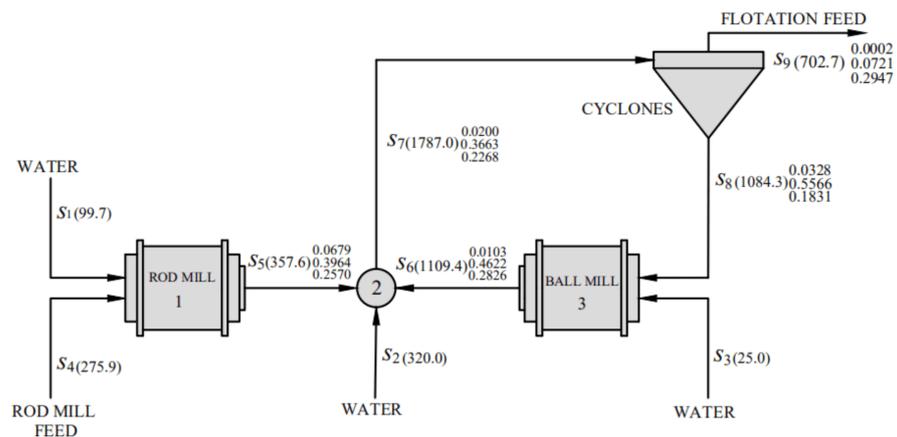


Figure 4. The metallurgical grinding flowsheet with true values of all the measured flowrates and mass weights of three components.

The third case, i.e., Pai-Fisher (PF) [34], is a typical nonlinear instance of data rectification, whose model is shown as the following Equation (20). The true values of the measured variables are $x = [4.5124; 5.5819; 1.9260; 1.4560; 4.8545]$ and true values of unmeasured variables are $v = [11.070; 0.61467; 2.0504]$. For all the measured variables, the corresponding σ of random noise is set as 2.5% of its true value.

$$\begin{cases} 0.5x_1^2 - 0.7x_2 + x_3v_1 + v_1v_2x_2^2 + 2x_3v_3^2 = 255.8 \\ x_1 - 2x_2 + 3x_1x_3 - 2x_2v_1 - x_2v_2v_3 = -111.2 \\ x_3v_1 - x_1 + 3x_2 + x_1v_2 - x_3v_3^{0.5} = 33.57 \\ x_4 - x_1 - x_2^2 + v_2 + 3v_3 = 0 \\ x_5 - 2x_3v_2v_3 = 0 \\ 2x_1 + x_2x_3v_1 + v_2 - v_3 = 126.6 \end{cases} \quad (20)$$

The fourth case, namely, the Swartz case (Sw) [35], is a heat exchanger network, where streams A_i ($i = 1, 2, \dots, 8$) is heated by streams B_i ($i = 1, 2, 3$), C_i ($i = 1, 2$) and D_i ($i = 1, 2$) via different heat exchangers, as Figure 5 shows. The true values of flowrate and temperature for each stream [12] are shown in Figure 5, too. The standard deviation of random noise for each flowrate is set as 2.5% of the corresponding true value of flowrate and 0.75 for temperature of each stream.

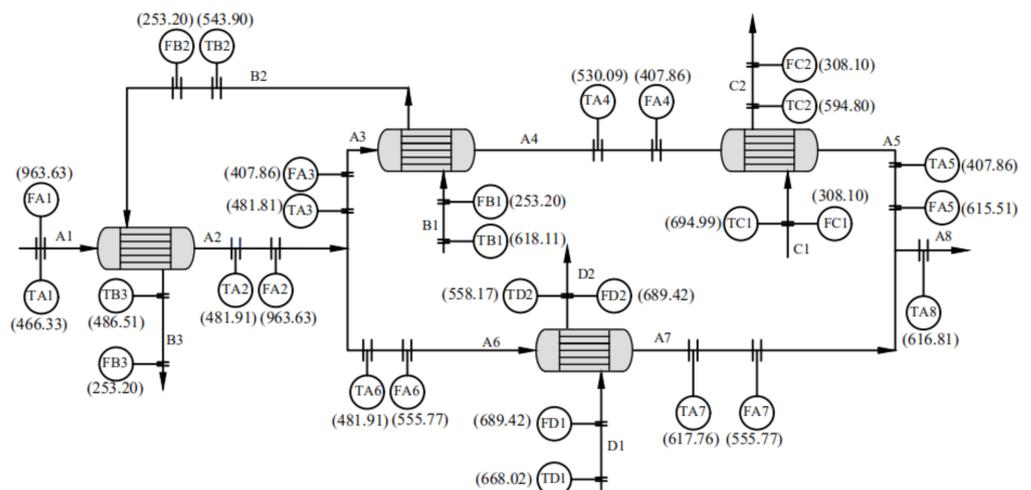


Figure 5. The Swartz flowsheet with true values of measured temperatures and flowrates.

For the Swartz case, both linear material balance equalities and nonlinear energy balance equalities for each heat exchanger/junction are used as constraints of data rectification. The enthalpy of unit mass of each stream is correlated with its temperature using a quadratic polynomial as Equation (21) shows, whose coefficients are shown in Table 2 [1].

$$H = v_1 + v_2T + v_3T^2 \quad (21)$$

Table 2. Coefficients of temperature-enthalpy correlation for each stream.

Stream	$A_i (i = 1, 2, \dots, 8)$	$B_i (i = 1, 2, 3)$	$C_i (i = 1, 2)$	$D_i (i = 1, 2)$
v_1	−6.8909	−14.8538	−28.2807	−11.4172
v_2	0.0991	0.1333	0.1385	0.1229
v_3	1.1081×10^{-4}	7.539×10^{-5}	9.043×10^{-5}	7.94×10^{-5}

For all the tested cases, the Monte Carlo simulations were carried out in a MATLAB 2018 (MathWorks, Boston, MA, USA) environment using a personal computer with Intel Core Processor (TM) i3 CPU 3120M @ 2.50 GHz, 8GB RAM (Intel, Santa Clara, CA, USA), random measuring noises were generated by “normrnd” command and “rand” command was used to assign the size and sign of a bias. The nonlinear programs of the EM were solved with “fmincon” command.

5. Results and Discussion

To evaluate the performances of different criteria of the modified EM approach, the OP and AVTI performances of the DC, PC and PDC for all the four tested cases are compared as shown in Figure 6.

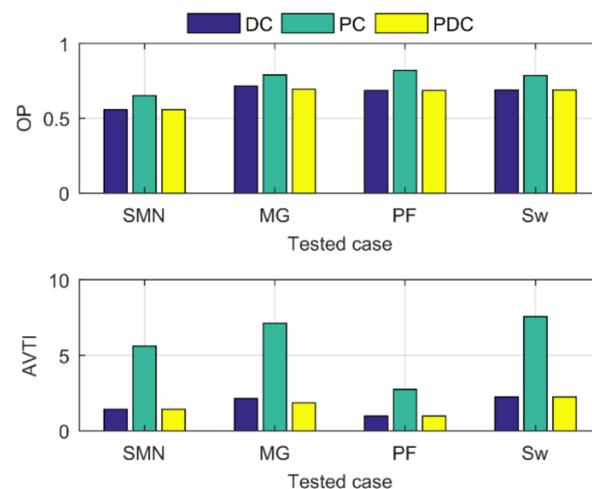


Figure 6. Performances of data rectification of bias detection criteria of modified EM approach. SMN: steam metering network; MG: metallurgical grinding; PF: Pai-Fisher; Sw: Swartz.

As Figure 6 shows, PC had higher OP and obviously higher AVTI than DC and PDC, which shows that the probabilities of random and gross error modes are not feasible to detect a bias because some variables not contaminated by a bias also have higher probability of gross error mode. The DC and PDC had the same OP and AVTI except for the bilinear MG case, where PDC detected a little less bias than DC; whether this was a special case needs to be investigated in the future, since this work focuses on modifying the original EM approach for rectifying measurements contaminated by persistent biases.

Because MEM-DC and MEM-PDC had almost the same performances of data rectification, MEM-DC was selected to be compared to the original EM approach, as shown in Table 3. As stated in Section 3.2, PC was used to detect bias for the original EM, because DC does not work in the situation of persistent bias contaminating measurement, the

standard deviation of random mode, i.e., $\sigma_{j,1}$, was enlarged by the bias contaminating the j th measurement, as shown in Figure 1, and DC based on the 3σ rule cannot detect any bias from experiences of the authors.

Table 3. The performances of the original and modified EM approaches for the tested cases.

Case	Total Biases	Method	OP	AVTI	RER	Time (s)
SMN	6984	EM	0.1668	3.0030	0.3542	8.3979
		MEM-DC	0.5488	1.4330	0.3628	1.0425
Grinding	6061	EM	0.2153	3.0980	0.4489	7.4701
		MEM-DC	0.6976	1.6590	0.4578	2.0415
Pai-Fisher	1213	EM	0.2003	0.9990	0.4804	0.6180
		MEM-DC	0.6735	1.0200	0.4838	0.5719
Swartz	2564	EM	0.1998	5.4910	0.7348	7.0294
		MEM-DC	0.6895	2.2480	0.7388	0.7423

As Table 3 shows, the original EM had much lower OP and much higher AVTI than MEM-DC, which shows that the persistent bias influences not only the standard deviation of random error mode, but also the probability of random and gross error modes. It is interesting that the original EM approach had only a little worse RER than MEM-DC, which shows that the rectified values of both approaches are close to each other. At last, MEM-DC obviously consumed much less time than the original EM, because fewer parameters needed to be estimated for the former one.

6. Conclusions

In this work, we analyze the influence of a persistent bias on the estimated standard deviation of the random error mode for the EM approach and argue that the 3σ rule cannot be used to detect bias under the occurrence of a persistent bias. A modified EM approach was devised by estimating the standard deviation of random error mode from process measurements before the EM iterations. The performances of the modified and original EM approaches were evaluated and compared through four widely used linear and nonlinear examples of data rectification, and the results show that the original EM approach cannot be used to detect persistent biases, while the modified EM can; the modified EM consumes much less time than the original EM due to the reduction of estimated parameters.

The convergence of the modified EM algorithm is not proved and we will study this in the future to increase our understanding of the EM approach and to increase the reliability of the proposed EM approach.

Author Contributions: Conceptualization, S.T.; methodology, W.J. and S.T.; software, W.J.; validation, R.L. and D.C.; formal analysis, W.J. and C.L.; investigation, R.L. and C.L.; resources, S.T.; data curation, W.J.; writing—original draft preparation, W.J. and S.T.; writing—review and editing, W.J., R.L., D.C. and C.L.; supervision, S.T.; project administration, W.J. and S.T.

Funding: This work is supported by the National Key Research and Development Program of China (No. 2019YFB2006300, No. 2019YFB2006305).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Valle, E.; Kalid, R.; Secchi, A.; Kiperstok, A. Collection of Benchmark Test Problems for Data Reconciliation and Gross Error Detection and Identification. *Comput. Chem. Eng.* **2018**, *111*, 134–148. [[CrossRef](#)]
2. de Menezes, D.Q.F.; de Sá, M.C.C.; Fontoura, T.B.; Anzai, T.K.; Diehl, F.C.; Thompson, P.H.; Pinto, J.C. Modeling of Spiral Wound Membranes for Gas Separations—Part II: Data Reconciliation for Online Monitoring. *Processes* **2020**, *8*, 1035. [[CrossRef](#)]
3. Qian, F.; Zhong, W.; Du, W. Fundamental Theories and Key Technologies for Smart and Optimal Manufacturing in the Process Industry. *Engineering* **2017**, *3*, 154–160. [[CrossRef](#)]
4. Crowe, C.M. Data Reconciliation—Progress and Challenges. *J. Proc. Cont.* **1996**, *6*, 89–98. [[CrossRef](#)]
5. Guo, S.; Liu, P.; Li, Z. Data Reconciliation for the Overall Thermal System of a Steam Turbine Power Plant. *Appl. Energy* **2016**, *165*, 1037–1051. [[CrossRef](#)]
6. Özyurt, D.; Pike, R. Theory and Practice of Simultaneous Data Reconciliation and Gross Error Detection for Chemical Processes. *Comput. Chem. Eng.* **2004**, *28*, 38–402. [[CrossRef](#)]
7. Llanos, C.; Sánchez, M.; Maronna, R. Classification of Systematic Measurement Errors within the Framework of Robust Data Reconciliation. *Ind. Eng. Chem. Res.* **2017**, *56*, 9617–9628. [[CrossRef](#)]
8. Soderstrom, T.; Himmelblau, D.; Edgar, T. A Mixed Integer Optimization Approach for Simultaneous Data Reconciliation and Identification of Measurement Bias. *Control. Eng. Pract.* **2001**, *9*, 869–876. [[CrossRef](#)]
9. Tao, S.; Zhang, D.; Xiang, X.; Jiang, W.; Li, R. Location Estimation Based MILP Approach for Multiple Gross Errors Identification. *Ind. Eng. Chem. Res.* **2019**, *58*, 18780–18787. [[CrossRef](#)]
10. Tao, S.; Xue, Y.; Xiang, S.; Jiang, W.; Li, R. Tighter Mixed-Integer Quadratic Programming Model for Process Data Rectification. *Ind. Eng. Chem. Res.* **2020**, *59*, 10061–10071. [[CrossRef](#)]
11. Câmara, M.M.; Soares, R.M.; Feital, T.; Anzai, T.K.; Diehl, F.C.; Thompson, P.H.; Pinto, J.C. Numerical Aspects of Data Reconciliation in Industrial Applications. *Processes* **2017**, *5*, 56. [[CrossRef](#)]
12. Rollins, D.; Cheng, Y.; Devanathan, S. Intelligent Selection of Hypothesis Tests to Enhance Gross Error Identification. *Comput. Chem. Eng.* **1996**, *20*, 517–530. [[CrossRef](#)]
13. 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. [[CrossRef](#)]
14. Arora, N.; Biegler, L.T. Redescending estimators for data reconciliation and parameter estimation. *Comput. Chem. Eng.* **2001**, *25*, 1585–1599. [[CrossRef](#)]
15. 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. [[CrossRef](#)]
16. Chen, J.; Peng, Y.; Munoz, J. Correntropy Estimator for Data Reconciliation. *Chem. Eng. Sci.* **2013**, *104*, 10019–10027. [[CrossRef](#)]
17. Llanos, C.; Sánchez, M.; Maronna, R. Robust Estimators for Data Reconciliation. *Ind. Eng. Chem. Res.* **2015**, *54*, 5096–5105. [[CrossRef](#)]
18. Yuan, Y.; Khatibisepehr, S.; Huang, B.; Li, Z. Bayesian method for simultaneous gross error detection and data reconciliation. *AIChE J.* **2015**, *61*, 3232–3248. [[CrossRef](#)]
19. Ng, S.K.; Krishnan, T.; McLachlan, G.J. The EM Algorithm. In *Handbook of Computational Statistics*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 139–172.
20. Karlis, D.; Evdokia, X. Choosing initial values for the EM algorithm for finite mixtures. *Comput. Stat. Data. Anal.* **2003**, *41*, 577–590. [[CrossRef](#)]
21. Karlis, D.; Ntzoufras, I. Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *J. Stat. Softw.* **2005**, *14*, 1–36. [[CrossRef](#)]
22. Fung, T.C.; Badescu, A.L.; Lin, X.S. A Class of Mixture of Experts Models for General Insurance: Application to Correlated Claim Frequencies. *ASTIN Bull.* **2019**, *49*, 647–688. [[CrossRef](#)]
23. Verbelen, R.; Gong, L.; Antonio, K.; Badescu, A. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bull.* **2015**, *45*, 729–758. [[CrossRef](#)]
24. Amin, F.; Choi, G.S. Hotspots Analysis Using Cyber-Physical-Social System for a Smart City. *IEEE Access* **2020**, *8*, 122197–122209. [[CrossRef](#)]
25. Amin, F.; Ahmad, A.; Choi, G. Towards Trust and Friendliness Approaches in the Social Internet of Things. *Appl. Sci.* **2019**, *9*, 166. [[CrossRef](#)]
26. Alighardashi, H.; Jan, N.M.; Huang, B. Expectation maximization approach for simultaneous gross error detection and data reconciliation using Gaussian mixture distribution. *Ind. Eng. Chem. Res.* **2017**, *56*, 14530–14544. [[CrossRef](#)]
27. Alighardashi, H.; Jan, N.M.; Huang, B. Data rectification for multiple operating modes A MAP framework. *Comput. Chem. Eng.* **2019**, *123*, 272–285. [[CrossRef](#)]
28. Johnston, L.; Kramer, M.A. Maximum likelihood data rectification: Steady-state systems. *AIChE J.* **1995**, *41*, 2415–2426. [[CrossRef](#)]
29. Do, C.B.; Batzoglou, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **2008**, *26*, 897–899. [[CrossRef](#)]
30. Morad, K.; Svrcak, W.Y.; McKay, I. A robust direct approach for calculating measurement error covariance matrix. *Comput. Chem. Eng.* **1999**, *23*, 889–897. [[CrossRef](#)]
31. Ripps, D.L. Adjustment of experimental data. *Chem. Eng. Prog. Symp. Ser.* **1965**, *61*, 8–13.

32. Serth, R.W.; Heenan, W.A. Gross error detection and data reconciliation in steam-metering systems. *AIChE J.* **1986**, *32*, 733–747. [[CrossRef](#)]
33. Serth, R.W.; Valero, C.M.; Heenan, W.A. Detection of gross errors in nonlinearly constrained data: A case study. *Chem. Eng. Commun.* **1987**, *51*, 89–104. [[CrossRef](#)]
34. Pai, D.C.C.; Fisher, G.D. Application of Broyden's method to reconciliation of nonlinear constrained data. *AIChE J.* **1988**, *34*, 873–876. [[CrossRef](#)]
35. Swartz, C.L.E. Data reconciliation for generalized flowsheet applications. In Proceedings of the Paper Presented at American Chemical Society National Meeting, Dallas, TX, USA, 16–20 March 1989.