Robust Data Reconciliation Based on a Generalized Objective Function

D. Wang and J. A. Romagnoli

Dept. of Chemical Engineering,
The University of Sydney, NSW 2006, Australia

Abstract: A robust data reconciliation method is proposed. This approach is based on the minimization of a generalized objective function, which is obtained from residuals probability density function estimation. Existing approaches are based on the minimization a pre-defined objective to take into account random and gross errors. The proposed approach is optimal in the sense of maximum likelihood estimation, allows the use of various model constraints and can be used for both steady state and dynamic systems. The performance of the proposed method is illustrated by a chemical engineering example. Copyright © 2002 IFAC

Key words: data reconciliation, robust estimation, robust statistics, gross errors, outliers detection

1. INTRODUCTION

A wide variety of process measurements are taken in chemical plants for various purposes such as process monitoring, control performance evaluation, process control and statistical quality control. Measurements can be contaminated with errors (random and/or gross) due to various sources such as measurement irreproducibility, instrument degradation and malfunction, human error, process-related errors, and other unmeasured errors (Crowe, 1996). Therefore we cannot expect that any set of measurements will obey the laws of conservation. Rational use of the large volume of data generated by chemical plants requires the application of suitable techniques to improve their accuracy. Data reconciliation is a procedure of optimally adjusting measured data so that the adjusted values obey the conservation laws and other constraints.

The conventional method for data reconciliation is the weighted least squares where adjustments to the data are weighted by the inverse of the measurement noise covariance matrix so that the model constraints are satisfied. The optimality/validity of this approach is based on a main assumption that the errors follow a normal Gaussian distribution. When this assumption is satisfied, conventional approaches provide unbiased estimates of the plant states. However, normal distribution usually does not exist in real chemical engineering practice, it is hard to assure the normality even for high quality measurements. The frequent presence of gross errors and outliers violates the assumptions in the conventional approach and makes the results invalid (Tamhane and Mah, 1985; Mah, 1990).

Several schemes have been suggested to deal with the corruption of conventional normal assumption. One of the usually adopted approaches is gross error and outliers detection, where the bulk of data is taken as normal distribution and small fraction of data which could not pass the statistical test is considered abnormal and has to be removed (Serth and Heenan, 1986; Narasimhan and Mah, 1987). This approach intends to take advantages of prevalence and convenience of normal assumption, but it has limitations. Firstly, the selected tests are based on the normality assumption, which is not necessarily true. The residuals from regression may be heavily biased by the presence of the outliers. Secondly, these gross errors detection approaches are for linear model. This is a restriction for almost any process model other than mass balances. Linearization is the most common approach for dealing with nonlinear constraints. However, by forcing the measurements to conform to the linearized (approximate) model, additional errors may be introduced. This may affect the validity of the tests.

The other approach is to take into account the non-ideality of the data distribution from the beginning, using an objective function that is constructed by maximum likelihood principles on contaminated error distributions (Tjoa and Biegler, 1991; Johnston...
and Kramer, 1995). In this case, the data reconciliation problem is posed again as an optimization problem and the error distribution contains two parts, a narrow Gaussian representing normal measurement noise and a broad Gaussian representing the gross errors. The two Gaussian distributions are weighted by the probability or not of having a gross error in a sensor respectively. The data reconciliation and gross errors detection can be performed simultaneously. There are no restrictions on the constraints, so it applies to both linear and nonlinear systems.

Recently, robust statistics ideas have been used to solve the problem of data reconciliation and parameters identification (Romagnoli and Sanchez, 2000; Albuquerque and Biegler, 1996.). This approach uses an objective function that, due to its mathematical nature, is insensitive to deviations from the ideal assumptions, especially to outliers. These estimators tend to look at the bulk of the data and ignore atypical values. In this fashion, an accurate regression can be performed even if nothing is known about the outliers or even the error structure of the data. The robust approach concentrates on finding estimators (the objective function in the optimization) whose influence functions are bounded. A number of objective functions can be chosen as candidates. Even through it is insensitive to outliers, the optimality of the estimation is still dependent on the distribution structure of the measurements errors according to maximum likelihood principles. If the errors of the bulk data do not follow the pre-defined distribution, corresponding to the robust objective function, the estimation will not be optimal.

In this article, an alternative robust DR approach is proposed for optimal data reconciliation in the MLE sense. Here the problem is posed as minimizing an objective function that reflects the real distributional structure of the measurement errors subject to the process model. The objective function is not pre-selected according to some robustness requirements, but obtained by estimation of the measurement error distribution. In this way it will retain the advantages of both robustness and maximum likelihood estimation.

The paper is organized as follows. The robust data reconciliation problem is introduced in the next section, and its features are discussed. The proposed robust DR approach based on a generalized objective function is then introduced and the estimation steps are described. An application example of steady-state DR with nonlinear constraints is provided to show the comparative advantages of the approach, followed by the conclusion of the work.

2. ROBUST DATA RECONCILIATION

Data reconciliation is usually posed as the following optimization problem,

\[ \min_{\hat{y}} \sum_{i=1}^{n} (y_{m_i} - \hat{y}_i)^{T} Q^{-1} (y_{m_i} - \hat{y}_i) \]

subject to

\[ f \left( \frac{d\hat{y}(t)}{dt}, \hat{y}(t) \right) = 0, \]

\[ h(\hat{y}(t)) = 0, \]

\[ g(\hat{y}(t)) \geq 0 \]

where, \( y_m, \hat{y}(t), f, h, g \) are discrete measurements, estimate functions, differential equation constraints, algebraic equality constraints, and inequality constraints, respectively. \( Q \) is the noise covariance matrix, \( Q = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2) \), \( \sigma \) is the measurement noise standard deviation vector. With the conventional approach one assumes that the measurement errors follow a certain (usually Gaussian) distribution, and all estimations and statistical inferences are based on that distribution. It has been shown that departures from the ideal distribution (such as gross errors and outliers) can invalidate these estimations and inferences. The use of robust estimators can alleviate this problem (Huber, 1981). In robust statistics, rather than assuming an ideal distribution, an estimator is constructed so that will give unbiased results in the presence of this ideal distribution, but also will be insensitive to deviation from ideality to a certain degree. The commonly used robust estimator is the M-estimator, which is a generalization of the maximum-likelihood estimator. Instead of using quadratic objective function, it uses an objective function, which reflects the error distribution structure in the optimization. Without changing the constraints, it is formulated as

\[ \min_{\hat{y}} \sum_{i=1}^{n} \rho(\varepsilon) \]

where, \( \rho(\varepsilon) = -\log f(\varepsilon), \varepsilon = y_{m_i} - \hat{y} \) is the error, \( f \) is the probability density function which considers the ideal distribution assumption as well as the deviation from the ideality. Usually, \( f \) is a contaminated distribution of random errors and gross errors/outliers,

\[ f = \left\{ f(\varepsilon) = (1 - \beta)N(\varepsilon | 0, \sigma^2) + \beta h(\varepsilon), 0 < \beta < 1 \right\} \]

where \( N(\varepsilon | 0, \sigma^2) \) is the zero-mean normal PDF with variance \( \sigma^2 \). The distribution of outliers is unknown but is assumed to belong to some class of symmetric distribution with zero-mean and finite variance. In particular, \( h(\varepsilon) \) is chosen to be any distribution characterizing the gross errors (e.g. normal PDF \( N(\cdot | 0, \sigma_g^2) \) with \( \sigma_g^2 \gg \sigma^2 \)) with \( \beta \) being the probable occurrence of outliers in the Gaussian distribution \( N(\cdot | 0, \sigma^2) \). \( \beta, \sigma_g \) are used as parameters in the data reconciliation to accommodate the distribution structure. In general, an estimator is said to be robust if a bounded shift from the ideal assumption will lead to a bounded shift in the estimates. The
in the objective function. For the least-squares estimator, the objective function is quadratic, that is \( \rho(\epsilon) = \frac{1}{2} \epsilon^2 \) and the influence function would be \( \psi_L(\epsilon) = \epsilon \). Which it is not robust, because the influence function will increase with respect to large errors. For the contaminated normal, the objective function is chosen as \( \rho(\epsilon) = -\log f(\epsilon) \).

\[
f(\epsilon) = (1 - \beta) \exp \left( -\frac{\epsilon^2}{2\sigma^2} \right) + \frac{\beta}{b} \exp \left( -\frac{\epsilon^2}{2b^2\sigma^2} \right)
\]

(4)

where, \( b > 1 \) is the ratio of the standard deviation of the gross-error distribution to the standard deviation of the random-error distribution. The influence function would be

\[
\psi_{\text{Cont}}(\epsilon) = \frac{(1 - \beta) \exp \left( -\frac{\epsilon^2}{2\sigma^2} \right) + \frac{\beta}{b} \exp \left( -\frac{\epsilon^2}{2b^2\sigma^2} \right)}{(1 - \beta) \exp \left( -\frac{\epsilon^2}{2\sigma^2} \right) + \frac{\beta}{b} \exp \left( -\frac{\epsilon^2}{2b^2\sigma^2} \right)}
\]

(5)

For very large values of \( \epsilon \), the weighting function \( w(\epsilon) \) can be approximated by \( w(\epsilon) \approx 1/b^2 \), where for small values of \( \epsilon \), \( w(\epsilon) \approx 1 \). Therefore, a good approximation to the influence function of the contaminated estimator would be

\[
\psi_{\text{Cont}} \approx \begin{cases} 
\frac{\epsilon}{b^2} & \epsilon \to 0 \\
1 & \epsilon \to \infty
\end{cases}
\]

(6)

This will down-weight the effect of the large errors. Even though its value is not unbounded when \( \epsilon \to \infty \), theoretically, it can be made bounded in a region by adjusting the parameter \( b \). The contaminated normal distribution has heavier tails than the pure normal distribution, making it less sensitive to outliers. In a similar way, one can use other distributions to describe the gross errors. If Laplacian distribution is used the objective function is given by:

\[
\rho(\epsilon) = \begin{cases} 
\frac{\epsilon^2}{2} & |\epsilon| \leq k \\
|\epsilon| - \frac{k^2}{2} & k < |\epsilon|
\end{cases}
\]

(7)

Thus, the influence function is

\[
\psi(\epsilon) = \begin{cases} 
-k & \epsilon < -k \\
\epsilon & -k \leq \epsilon \leq k \\
k & k < \epsilon
\end{cases}
\]

(8)

where, \( k \) is the parameter characterizing the contamination degree (Wang, et. al 2000). This influence function is bounded for any value of the errors.

According to the robustness requirement, one can just design an objective function without taking into consideration the corresponding error distribution structure, such that its influence function is bounded. The fair function, for example, can be used in the optimization (Albuquerque and Biegler, 1996)

\[
\rho(\epsilon) = c^2 \left( \frac{\epsilon}{c} - \log(1 + \frac{\epsilon}{c}) \right)
\]

(9)

The influence function is in this case

\[
\psi(\epsilon) = \frac{\epsilon}{1 + \frac{\epsilon}{c}}
\]

(10)

which is bounded with the argument \( \epsilon \) approaching to infinity. The influence curves of the above estimators are shown in Figure 1.

Even through the above approaches are less sensitive to the gross errors and outliers, the optimality of the estimation is still dependent on the suitability of the chosen function with respect to the actual distribution of the data. It is generally hard to characterize the distribution of the errors correctly. If the real errors do not follow the specified distributions, the estimation could be biased. For example, in Eq. (3) the contamination degree \( \beta \) needs to be predicted a priori, and in (9) special attention is needed on the choice of \( c \) in the objective function.

3. ROBUST DATA RECONCILIATION BASED ON A GENERALIZED OBJECTIVE FUNCTION

3.1. The method

A generalized objective function can be used in the optimization for data reconciliation,
where, \( \rho(\varepsilon) = -\log f(\varepsilon) \), \( \varepsilon = y_m - \hat{y} \) is the error. In comparison to previous approaches, where the error probability density function \( f \) is constructed in advance considering the ideal distribution assumption as well as the deviation from the ideality, here \( f \) is estimated from the real residual data. Thus, the objective function will accommodate for the real error distribution and the estimator will be optimal in the MLE sense. In general this will result in an iterative estimation procedure with an increase in the computational load. In order to speed up convergence and improve the performance, one of the previous approaches can be employed in a preliminary step during the DR procedure. The basic steps of the proposed robust data reconciliation procedures are:

1. Find a reasonable robust preliminary estimator that is not greatly influenced by the outlying observations and take the estimation results as the initial values.

2. Calculate the residuals from this preliminary fit. Based on these residuals and using the PDF estimation technique, determine the objective function, which is optimal for these residuals.

3. Using the objective function (constructed in (2)) with the preliminary estimates from (1) as starting values, perform one-step iteration toward the M-estimator.

4. If the residuals converge or satisfy the pre-specified tolerance, the procedure is terminated. Otherwise, go to (2).

Figure 2 gives the flowchart of the proposed robust data reconciliation based on a generalised objective function.

In practice, we have a set of observed data points that are assumed to be samples from an unknown probability density function. Density estimation is the construction of an estimate of the density function from the observed data. Techniques to estimate the density function of a given data set are classified into two categories: parametric approach and non-parametric approach. The former assumes that the data belong to one of a known family of distributions and then regresses any required parameters. When there is a prior knowledge that the underlying density function \( f \) is of a certain type, then the appropriate parametric technique can be used. Unfortunately, for most processes the underlying distribution of the data is not known and most likely does not follow a particular class of density function. Therefore, one has to estimate the density function using an unstructured approach. This is called non-parametric estimation.

Among most of commonly used methods, kernel density estimation method has been chosen for our study due to its wide applicability, smoothness and simplicity. Kernel estimate with kernel \( K \) is defined by

\[
\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)
\]

where, \( h \) is the window width, also called the smoothing parameter or bandwidth. The quality of an estimate is now widely recognized to be primarily determined by the choice of the smoothing parameter \( h \), and only in a minor way by the choice of kernel \( K \). For simplification, kernel \( K \) usually, but not always, will be a symmetric probability density function, e.g. the normal density.

\[
\min \sum \rho(\varepsilon)
\]

(11)
The problem of choosing how much to smooth is of critical importance in density estimation. A number of alternative measures exist to estimate \( h \) and can be found in Scott (Scott, 1992). The appropriate choice for the smoothing parameter is, in fact, influenced by the purpose for which the density estimate is to be used. In this work, the smoothing parameter is selected by least squares cross-validation (Silverman, 1986). It is completely automatic and data driven.

Given any estimate \( \hat{f} \) of a density \( f \), the integrated square error can be written

\[
\int \left( \hat{f} - f \right)^2 = \int \hat{f}^2 - 2\hat{f}f + f^2
\]

The last term of (14), does not depend on \( \hat{f} \), and so the ideal choice of window width will correspond to the choice, which minimizes the first two terms. The basic principle of least-squares cross validation is to construct an estimate of \( \hat{f} \) from the data themselves and then to minimize this estimate over \( h \) to give the choice of window width. Detailed discussion can be found in literature (Silverman, 1986).

4. SIMULATION CASE STUDY

The heat-exchanger network example, introduced originally by Swartz (1989), will be used to demonstrate the performance of the proposed robust data reconciliation method (Figure 3).

Process stream \( A \) is heated by process streams \( B, C \) and \( D \) at various junctions. The standard deviations of flow rates and temperatures are 2% and 0.75\textdegree Celsius respectively. The system has 16 measured variables, 14 unmeasured variables and 17 constraints, which describe total mass and energy balances around the exchangers, mixers, and dividers (Romagnoli and Sanchez, 2000). Several sets of measurements were generated for our study by adding the simulated values with different type error distributions such as: Gaussian, uniform, \( t \), non-central \( t \) distribution. The presence of outliers was also generated randomly for each set accordingly. 80 sampling times were specified in each set. Three estimation methods (least squares objective function, tailored objective function based on bivariate likelihood distribution and generalized objective function based on distribution estimation) are used for each set of data to compare their relative performances.

The performance criterion used to compare the efficiencies of the various estimation methods was the mean-squared error (MSE):

\[
MSE = \frac{1}{nK} \sum_{j=1}^{K} \sum_{i=1}^{n} \left( \tilde{y}_{i,j} - y_{i,j} \right)^2 \sigma_n^2
\]

Where, \( n \) is the dimension of the measurement vector, \( K \) is the number of measurement vectors used to calculate the MSE, \( n=16, K=80 \) in this study). \( y_{i,j} \) is the true plant state, \( \tilde{y}_{i,j} \) is the rectified state, and \( \sigma_n^2 \) is the standard deviation of the Gaussian noise on sensor \( n \). The MSE is constructed such that better performance is obtained if its value is driven toward zero.

The comparative results are listed in Table 1. It can be seen for the table that for normal noise all three methods have similar performance. However, with outliers occurring in the measurements, the performance of the least squares method deteriorates considerably, while the robust approaches reveal their distinctive advantages. The estimates based on the generalized method shows always better performance than the method based on a pre-determined error distribution. This is especially important for the cases where the noises are generated by distributions other than the contaminated Gaussian. The reason lies in that the objective function used is not appropriate or optimal to represent the real error’s distribution or the changes in the error’s distribution. On the other hand

<table>
<thead>
<tr>
<th></th>
<th>Least Squares</th>
<th>Tailored Objective Function</th>
<th>Generalized Objective Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>6.78</td>
<td>6.91</td>
<td>6.86</td>
</tr>
<tr>
<td>(with outliers)</td>
<td>12.33</td>
<td>8.41</td>
<td>8.44</td>
</tr>
<tr>
<td>Uniform</td>
<td>8.58</td>
<td>6.75</td>
<td>6.54</td>
</tr>
<tr>
<td>(with outliers)</td>
<td>16.58</td>
<td>8.72</td>
<td>8.21</td>
</tr>
<tr>
<td>T</td>
<td>12.33</td>
<td>5.75</td>
<td>4.38</td>
</tr>
<tr>
<td>(with outliers)</td>
<td>17.11</td>
<td>5.34</td>
<td>4.19</td>
</tr>
<tr>
<td>Non-central T</td>
<td>23.47</td>
<td>14.58</td>
<td>6.79</td>
</tr>
<tr>
<td>(with outliers)</td>
<td>29.31</td>
<td>17.63</td>
<td>5.34</td>
</tr>
</tbody>
</table>
the generalized objective function can accommodate for the changes and thus adapting to the new conditions. Figure 4 shows the comparison between the different estimation methods at one sensor (flowrate A1). The measurements were generated by adding to the simulated values a non-central \( t \) distribution noise and outliers.

5. CONCLUSION

Measurements taken from process systems often contain gross errors, outliers as well as random errors, which make it hard to specify the error structure in advance. Robust approaches can alleviate the problem by designing estimators, which are insensitive to the deviation from ideality. However, the existing robust approaches lack the optimal performance in maximum likelihood sense because the objective function used in optimization does not reflect the actual data distribution. An alternative robust data reconciliation procedure based on a generalized objective function has been introduced and its advantages demonstrated. The proposed approach is optimal in the MLE sense while keeping its robustness features. Follows the real data distribution, allows the incorporation of various model constraints and it can be used for both steady state and dynamic systems. Simulation studies show better performance when compared with existing approaches.

6. REFERENCES


