Wavelet-Based Adaptive Robust M-Estimator for Nonlinear System Identification

D. Wang and J. A. Romagnoli
Process System Engineering Laboratory, Dept. of Chemical Engineering, The University of Sydney, NSW 2006, Australia

A. A. Safavi
Dept. of Control Engineering, Faculty of Electronic Industry, Shiraz University, Shiraz, Iran

A wavelet-based robust M-estimation method for the identification of nonlinear systems is proposed. Because it is not based on the assumption that there is the class of error distribution, it takes a flexible, nonparametric approach and has the advantage of directly estimating the error distribution from the data. This M-estimator is optimal over any error distribution in the sense of maximum likelihood estimation. A Monte-Carlo study on a nonlinear chemical engineering example was used to compare the results with various previously utilized methods.

Introduction

Robust system identification has been an important research area during the 1990s in an attempt to achieve more advanced and robust control. However, it is still in its initial stage compared with classical system identification methods (Wu and Cinar, 1996). In the classical system identification techniques the parameters of the system are obtained by minimizing an objective function that reflects the distribution of the error subject to the process model under the *a priori* assumptions. The combination of normally distributed errors and ordinary least-squares (OLS) estimates has prevailed. The optimality of OLS is conditional upon the properties of the errors. If the errors are independent and identically distributed (IID) and normal, we will get an unbiased and efficient parameter estimation of the system in the sense of maximum likelihood estimation (MLE).

However, the problem arises when the errors do not satisfy the *a priori* assumption. If there are some outliers in the data (*a posteriori* information), the error distribution is no longer normal and the estimates will be biased (Albuquerque and Biegler, 1996). To deal with this problem, outlier detection/elimination techniques have been suggested which are based on some preselected rules to keep the data within a normal region (Narasimhan and Mah, 1988; Tamhane et al., 1992). In fact, the error distribution is not always normal and simply deletion of the data may cause partial loss of information about the system (Hampel et al., 1986). Furthermore, this technique will add computational load for on-line identification (Albuquerque and Biegler, 1996).

Robust system identification methods are needed to alleviate these problems by constructing an estimator that will give unbiased estimates when the error follows *a previously known* distribution and still behaves well if there are deviations from ideality. Furthermore, robust process control of a system requires a robust system identification method that can provide optimal estimates for a broad class of distributions or any distribution of errors regardless of the effect of kurtosis and skewness.

Several robust system identification methods have been considered in the literature. One approach is the use of L_p estimate in which we estimate the parameters by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i,\theta)|^p$$

instead of the sum of errors square (Butler et al., 1990). Least absolute deviation (LAD) and OLS estimation correspond to $p=1$ and $p=2$, respectively. Values of p less than two are associated with tails which are thicker than the normal distribution. Forsayth (1972) suggests that $p=1.5$ can give good

Correspondence concerning this article should be addressed to J. A. Romagnoli.

robustness features for a thicker tailed distribution. L_p estimators can be regarded as, in fact, special cases of robust M-estimates, because $p = 1$, 2 correspond to Laplace and Normal error distributions, respectively (Poljak and Tsypkin, 1980; Butler et al., 1990).

Huber's M-estimates (Huber, 1981) provide the general framework for robust estimation. The idea is straightforward in the sense of MLE, where the performance index for the least favorable distribution within a given class is minimized. Hampel et al. (1986) developed the idea and gave the concept of Influence Function (IF) to describe the IF's properties and its effects on robust estimation. The crucial step in M-estimation for robust system identification is the choice of these influence functions. Based on different pre-selected choices of the influence functions, we get different robust estimators. In fact, these functions are determined by the error distribution of the system model. Butler et al. (1990) gave the result based on the generalized t distribution, which includes as special cases OLS, LAD, and L_p, as well as some estimation procedures which have bounded and redescending influence functions. However, these functions are just suited for the case when the underlying distribution is symmetric. To deal with skewed and heavy-tailed distribution, Moberg et al. (1980) and Yuh and Hogg (1988) suggested a method of choosing the influence function according to distribution classification based on a general lambda distribution. However, the disadvantage of this technique appears when the distribution is quite different from the class of the distribution such as multi-modal distribution, or when the distribution is close to the classification bounds. In this case, the estimates are far from optimal.

To improve this, Wu and Cinar (1996) use polynomials to approximate the influence function based on the generalized exponential distribution which includes previous distributions (Wu and Cinar, 1996; Lye et al., 1993). However, this method requires that the order of the polynomials be predetermined for approximation and also to assume $d(u) = 1$ (one type of the error probability density function (pdf), Cobb et al. (1983)) which in turn limits the range of the error distribution (Wu and Cinar, 1996).

In this article, an alternative approach is proposed for robust system identification based on wavelet estimation of the density function. This method is not based on any generalized broad class error distribution assumption, so it is a flexible and nonparametric approach. In addition, it can deal with the dependent observations (Safavi et al., 1997).

The robust system identification problem is introduced, and then several robust techniques and their features are discussed. Secondly, the proposed robust estimation steps and the wavelet method for robust estimation are introduced. Some advantages of the wavelet estimation approach are also discussed. A Monte-Carlo study on a nonlinear chemical engineering example is provided to compare several robust estimation methods, followed by the conclusions of this work.

Robust M-Estimation for Nonlinear System Identification

To grasp the essence of the robust system identification method, let us consider a nonlinear system described by the following model

$$y = f(x, \theta) + e(t) \tag{1}$$

where y is the output, x is the input, $e(t)$ is noise, and θ is the k dimension system parameters to be identified. The OLS estimate of parameters vector θ is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(x_i, \theta)\right]^2 \tag{2}$$

or, equivalently, by solving the system of k equations which results from differentiating Expression 2 with respect to θ and setting the derivatives equal to zero

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\left[y_i - f(x_i, \theta)\right]^2 = 0 \quad j = 1, 2, \ldots, k \tag{3}$$

If the errors are normally distributed, the OLS estimate is the maximum likelihood estimate of system parameter vector θ, and the estimate is unbiased and efficient (minimum variance estimates) in the statistical sense. However, if there are outliers in the data, the underlying distribution is not normal, and, thus, the OLS will be biased. Huber has demonstrated the inefficiency of OLS regression estimates if the underlying errors are not normally distributed and even a single grossly outlying observation will damage the estimates (Huber, 1972, 1981).

To solve this problem, a more robust method of estimation is needed. Since the true underlying distribution is not always normally distributed (and never known exactly), it seems sensible to use an estimation method which works well for the case that the underlying distribution has some deviations from the ideal one, and has application in a wide variety of possible situations. Huber (1981) discussed some robust procedures and generated the M-estimator approach for robust estimation of parameters θ in Expression 1. In this approach the quadratic implicit function in Expression 2 is replaced by a general ρ function.

Denoting the residual values by

$$u_i(\theta) = y_i - f(x_i, \theta) \quad i = 1, 2, \ldots, n \tag{4}$$

the M-estimates of vector θ is given by

$$\hat{\theta} = \arg\min\frac{1}{n}\sum_{i=1}^{n}\rho\left[u_i(\theta)\right] \tag{5}$$

or

$$\hat{\theta} = \arg\min\frac{1}{n}\sum_{i=1}^{n}\rho\left[y_i - f(x_i, \theta)\right] \tag{6}$$

where ρ is usually a convex function in order to ensure the solution to be unique.

In contrast to Expression 3 the estimate can be defined as the solution of the k equations

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left[y_i - f(x_i, \theta)\right]\cdot\frac{\partial f(x_i, \theta)}{\partial\theta_j} = 0 \quad j = 1, 2, \ldots, k \tag{7}$$

where $\psi(u) = \rho'(u)$. If ρ is convex and ψ is continuous, the definitions in Expressions 6 and 7 are equivalent. For all but the simplest ψ functions, the solution to Expression 7 must be obtained numerically, generally by using some iterative approximation processes.

Occasionally, it will be convenient to refer to the ρ function in Expression 6, but generally, the form Expression 7 will be used in the robust M-estimation approach. The use of the ψ form is due to Hampel's concept of influence function (IF) (Hampel et al., 1986). Hampel gave the infinitesimal approach of the residuals and derived the ψ to be proportional to the IF which reflects the influence of a single residual on the estimation. According to the IF concept, the values of ψ represent the effect of the residuals on the parameter estimation. If ψ is unbounded, it means that an outlier has an infinite effect on the estimation. Thus, the most important requirement for robustness is that ψ must be bounded and it should have a small value when the residual is large. In fact, the value of the ψ function corresponds to the gross-error sensitivity (Hampel et al., 1986); it measures the worst (approximate) influence that a small amount of contamination of fixed size can have on the value of the estimate.

Huber's approach is an inspiration for robust system identification, and Hampel's influence function is heuristic for the determination of the ψ function (where $\psi = \rho'$, as we mentioned above). For an estimator to be robust, its influence function, which is characterized by ψ, has to be bounded as the observation u is taken to infinity, no matter which error distribution u belongs to.

For least-squares estimation, we take the error square function as ρ, and the influence function $\psi = u$. Taking a very large observation u arbitrarily, $u \to \infty$, ψ will grow to infinity, and this means a single outlier has a big influence on the estimation. From this viewpoint, in the least-squares estimation every observation will be treated equally and has the same weight. Figure 1 shows the ψ function for least-squares estimation, where ψ is proportional to observation u.

For robust identification, we now turn to the problem of how to properly choose the ψ function which will meet the crucial condition that it is bounded when the argument u is very large, while at the same time ψ must be continuous. In other words, we do not treat the observation equally as least-squares does, we give a limited (usually small) weight for very large error observation, and we may also give zero weight to ignore its contribution to the estimation.

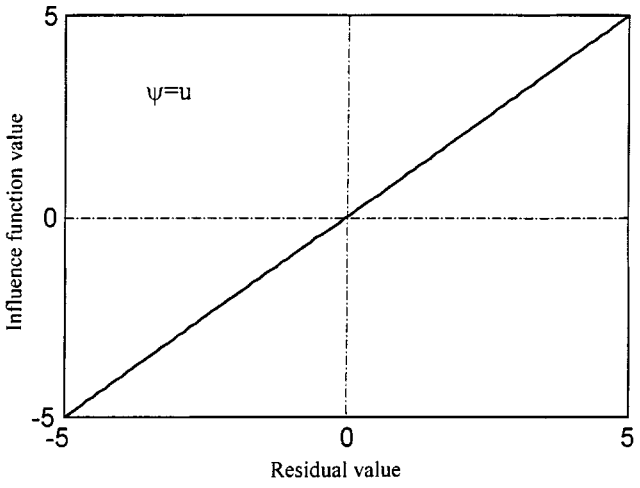

Figure 1. Influence function of least-square estimator ($\Psi \to \infty$ as $u \to \infty$).

The following are some choices of ψ functions which downweight or omit extreme values

Huber's:

$$\psi_h(u) = \min\{b, \max[u, -b]\} = u \cdot \min\left\{1, \frac{b}{|u|}\right\} \quad (8)$$

where b is a pre-selected value which is the bound of weight.

Hampel's:

$$\psi(u) = \begin{cases} u & 0 \le |u| \le a \\ a \cdot \text{sign}(u) & a \le |u \le b| \\ a \cdot \dfrac{r - |u|}{r - b} \text{sign}(u) & b \le |u| \le r \\ 0 & r \le |u \end{cases} \quad (9)$$

Andrews':

$$\psi(u) = \begin{cases} \sin\left(\dfrac{u}{c}\right) & |u| \le c\pi \\ 0 & |u| \ge c\pi \end{cases} \quad (10)$$

Tukey's:

$$\psi(u) = \begin{cases} u\left(1 - \dfrac{u^2}{k^2}\right)^2 & |u| \le k \\ k\,\text{sign}(u) & |u| > k \end{cases} \quad (11)$$

Fair Function (Albuquerque and Biegler, 1996)

$$\rho(u) = c^2\left[\frac{|u|}{c} - \log\left(1 - \frac{|u|}{c}\right)\right]$$

$$\psi(u) = \frac{u}{1 + \dfrac{|u|}{c}} \quad (12)$$

For the case where the underlying distribution is light-tailed or skewed, Moberg et al. (1980) use the residuals from the preliminary fit to classify the error distribution according to tail-weight and skewness characteristics. Moberg et al. used the classification to select a ψ function from a set of five such functions constructed so that each individual function is effective for a specific class of distribution (such as the ones that are light-tailed and symmetric) and so that collectively they cover a broad class of possible distribution.

The general form of the influence function for skewed and light-tailed distribution by Moberg is Moberg's

$$\psi(u) = \frac{a^2 u}{(b + u)^2 + c^2} \quad (13)$$

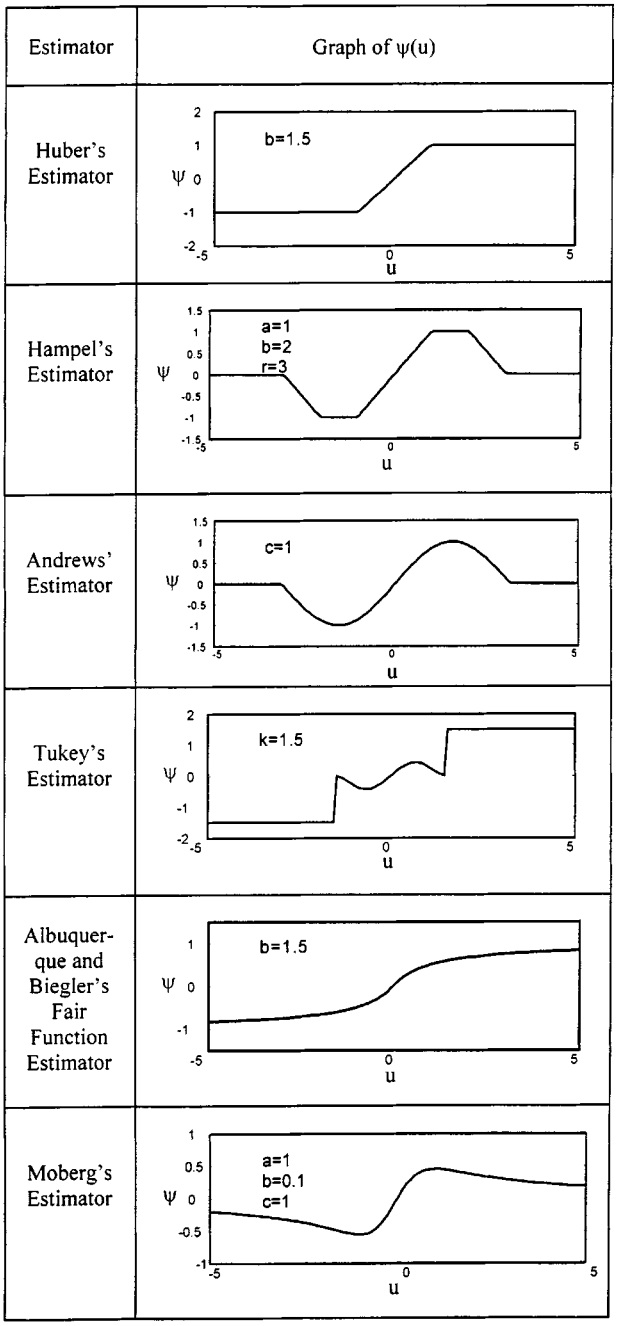

Figure 2. Examples of robust M-estimators.

Figure 2 shows these examples of M-estimators. All the ψ functions are pre-selected in order to downweight or omit the influence of extreme values. All the examples (except Moberg's) are skew-symmetric (that is, $\psi(u) = -\psi(-u)$) that are basically intended to be used with symmetric error distributions. For Moberg's estimation, the general form of the ψ function is predetermined according to the IF strategy, and the parameters a, b, and c are determined according to the classification of the error observation.

In fact, the M-estimates of the parameters in Expression 1 are defined in terms of real value ψ functions as described in Expression 7. For MLE estimation of the parameters, the ψ function can be chosen as follows

$$\psi(u) = -\frac{p'(u)}{p(u)} \tag{14}$$

where p is the density function of the true underlying distribution $\{u(t)\}$. This is a maximum likelihood estimation of parameters vector θ and it possesses the smallest asymptotic variance $V(\psi, F)$, which is the inverse of the Fisher information (Hampel et al., 1986).

From Expression 14, we find that choosing the influence function is equivalent to obtaining the density function of $\{u(t)\}$. If the error density function can be obtained, the robust M-estimation problem can be solved. This is the reason why so many authors approach the M-estimation based on different generalized error distributions. However, these generalized distributions approaches will limit the range of error distribution because we have to assume at first that the error belongs to that distribution while frequently it does not. We believe that the most reasonable approach should be based on the residuals of the data rather than pre-selection of a distribution. This seems to be a more realistic approach.

Wu and Cinar (1996) use a polynomial approximation of errors of density function p based on a generalized exponential family given by

$$p(u) = \xi \exp\left[-\int^{u} \frac{g(v)}{d(v)} dv\right] \tag{16}$$

where g and d are reasonable functions and ξ is scale constant. Generally, g and d are polynomials, but they can be allowed to include more general functions such as logarithms, trigonometric expressions, and so forth. On differentiation with respect to u, Expression 16 yields

$$\frac{p'(u)}{p(u)} = -\frac{g(u)}{d(u)} \tag{17}$$

and then Expression 17 is used for the calculation of ψ. Detailed explanation can be found in Wu and Cinar (1996) and Cobb et al. (1983). The disadvantage of this approach is the need to pre-determine the order of the polynomials.

Wavelet Framework of Robust Estimation

In the following we first describe the wavelet-based robust identification procedure, and then the features of wavelet density estimators is discussed.

Wavelet-based Adaptive Robust M-Estimator (WARME)

There are obvious advantages in using wavelet approach to construct the ψ function:

(1) It is not necessary to assume that the underlying errors belong to some generalized distribution as other approaches do. It is, in fact, difficult to state that a data set exactly belongs to a certain class of distribution. Although parametric methods are widely used, the drawbacks are obvious when the true data do not fit the distribution well. The nonparametric approaches often have more practical meaning.

(2) Pre-selection of any form of the ψ function is not required; the ψ function can be obtained directly from the residual data, so it is adaptive and flexible.

(3) It is straightforward to obtain the ψ function using the wavelet approach as will be discussed later.

The basic steps of wavelet based robust parameter M-estimation 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 from Step 1. Based on these residuals, using the wavelet estimation technique determine the ψ function which is optimal for these residuals.

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

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

Figure 3 gives the flowchart of the wavelet-based robust parameter M-estimation procedure.

The development of wavelet-based robust M-estimates involves selecting preliminary estimates to construct the ψ function and solving implicitly simultaneous equations. The most commonly suggested starting values for iterative procedures are the L_p estimates. Forsayth (1972) used OLS (L_2) starting values in finding L_p estimates ($1 < p < 2$). Andrews (1974) developed a procedure for finding starting values which are somewhat analogous to the sample median, but their procedure requires considerable computation. Other more robust estimators which have high breakdown points can be used as the starting estimators (Stromberg and Ruppert, 1992). Hill and Holland (1974) found that L estimates provided good starting values for M-estimates computations, and we follow their recommendation. Constructing the optimal ψ function in Step 2 above is the key to the robust M-estimation procedure. The function constructed should satisfy the error distribution. The complete wavelet-based approach of density estimation will be discussed in the following subsections.

In the general case the nonlinear equation in Expression 7 cannot be solved explicitly, and to find θ iterative methods should be employed whereby a sequence $\theta_1, \theta_2 \ldots, \theta_m, \ldots$ is composed which converges to θ. The simplest method is the gradient

$$\theta_{m+1} = \theta_m - \gamma_m \sum_{i=1}^{n} \psi(u_i^m) \nabla_\theta f(x_i, \theta_m) [u_i^m = f(x_i, \theta_m) - y_i] \tag{18}$$

In the smooth case $\psi(u)$ and $f(x, \theta)$ are continuously differentiable, γ_m can be selected constant ($\gamma_m \equiv \gamma$, $\gamma > 0$ is sufficiently small), or so as to satisfy the steepest descent condition $\gamma_m = \arg \min \sum_{i=1}^{n} \rho[\theta_m - \gamma \sum_{i=1}^{n} \psi(u_l^m) \nabla_\theta f(x_i, \theta_m)]$. Gradient type methods with fast convergence (such as conjugate gradient or the method of variable metric) can be employed. If $\rho(u)$ is nonsmooth convex, then $\psi(u)$ should be understood as a sub-gradient and γ_m should satisfy the condition $\gamma_m \to 0$, $\Sigma\gamma_m \to \infty$.

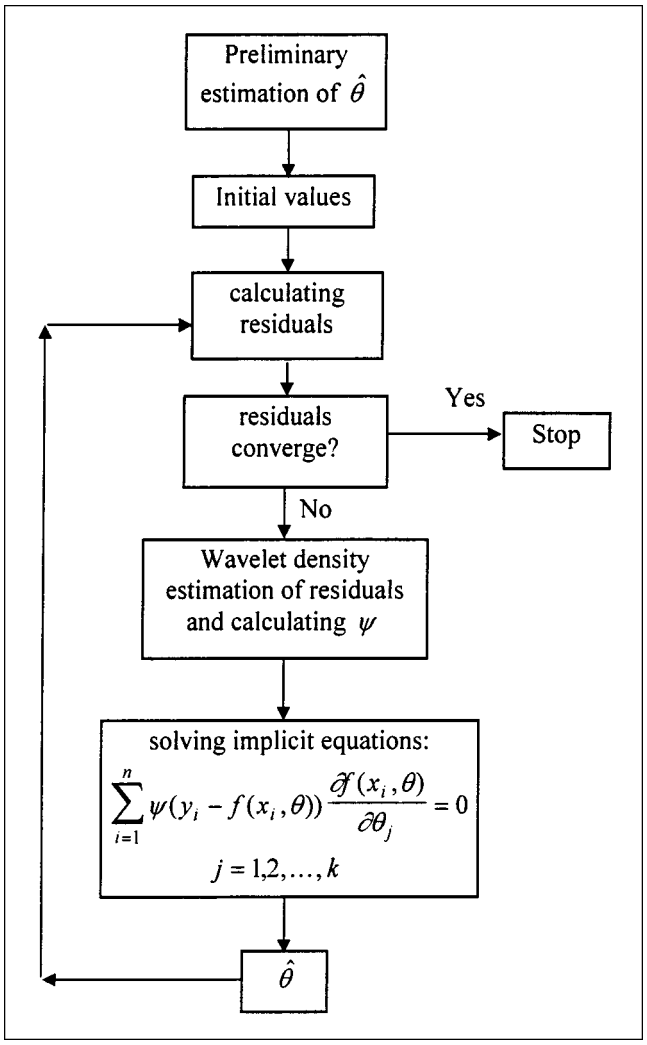

Figure 3. Procedure for wavelet-based adaptive robust M-estimation.

For smooth problems, Guass-Newton methods can work and these method can be shown to converge with probability 1 fast enough under certain assumptions for a good initial approximation θ_0. For a nonsmooth problem, its convergence can also be proved. Detailed discussion on convergence can be seen in Poljak and Tsypkin (1980).

The procedure has been implemented within a PC-MATLAB environment and the application results are presented in the fourth section.

Probability density function estimation using wavelets

Wavelet-based density estimators are extensively discussed by Safavi et al. (1997). Here, a brief discussion is presented. Wavelet estimators belong to the family of orthogonal estimators. The idea of using an orthogonal basis in estimating a density function was first considered by Cencov (1962). The orthogonal basis estimators approach the density estimation

problem from quite a different point of view, although some may find some similarities between this approach and the kernel method. The idea, here, is to estimate p by estimating the coefficients of its orthogonal projections on a basis. Recently, with the introduction of wavelet orthonormal bases (see Mallat, 1989; Daubechies, 1992) an increasing attention has been given to the use of these new bases for the estimation of the density function, for example, Kerkyacharian and Picard (1992).

Wavelets are a new family of localized basis functions and have found many applications in large areas of science and engineering (see Safavi, 1996). These basis functions can be used to express and approximate other functions. They are functions with a combination of powerful features, such as orthonormality, locality in time and frequency domains, different degrees of smoothness, fast implementations, and in some cases compact support. Wavelets are usually introduced in a multiresolution framework developed by Mallat (1989).

Let $\{x\}_{i=0}^{n-1}$ be a stationary process and p the density function of x. Then, invoking wavelets and multiresolution analysis, the approximation of p at resolution m is p_m

$$p_m(x) = \sum_{k=-\infty}^{\infty} a_{m,k}\phi_{m,k}(x) \tag{19}$$

or, in general, $p(x)$ is expressed as (see previous arguments)

$$p(x) = \sum_k a_{m,k}\phi_{m,k}(x) + \sum_{l=-\infty}^{m}\sum_k d_{l,k}\psi_{l,k}(x) \quad k \in Z \tag{20}$$

with

$$a_{m,k} = \langle \phi_{m,k}, p \rangle$$
$$d_{l,k} = \langle \psi_{l,k}, p \rangle \tag{21}$$

where ϕ_s and ψ_s are, respectively, the scaling functions and wavelets in a multiresolution framework. Since p is not available, the unbiased estimation of each coefficient in Eq. 21 can be computed as

$$\hat{a}_{m,k} = \frac{1}{n}\sum_{i=0}^{n-1}\phi_{m,k}(X_i)$$
$$d_{l,k} = \frac{1}{n}\sum_{i=0}^{n-1}\psi_{l,k}(X_i) \tag{22}$$

Having obtained some illustration of wavelet-based density estimators, some advantages of this type of estimator are briefly explained here. The nature of wavelets and multiresolution analysis makes the wavelet estimators superior to the other estimators in general. Nevertheless, one may emphasize on the following points:

(1) The ease of computation of this estimator is very attractive.

(2) Local learning and adaptation is reasonably well achieved with wavelet estimators.

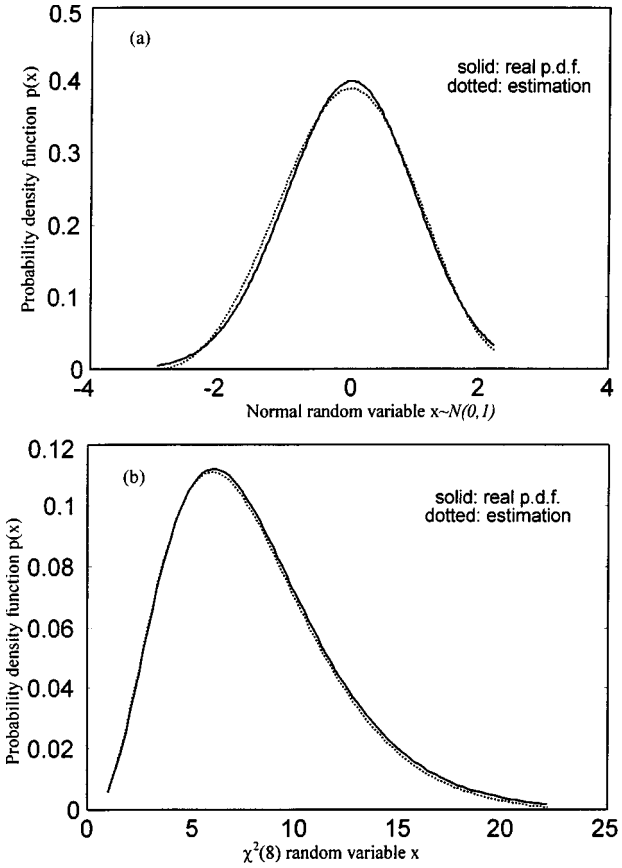

Figure 4. Density estimation of random variable with wavelets.

Horizontal axes are the arguments; the vertical axes are probability density functions as indicated. Number of data: $n = 100$.

(3) The multiresolution representation of the estimation allows the user to choose the degree of accuracy against the simplicity (or the speed) of estimations.

(4) While other estimators fail to deal with dependent observations, wavelet estimators challenge dependent observations in several ways.

Figure 4 presents the comparative results of wavelets based density estimation. Detailed discussion on wavelet estimators and their advantages over other estimators is presented in Safavi et al. (1997).

Case Study

In this section we present a Monte-Carlo study comparing four parameter estimation procedures using different error distributions contaminated by isolated outliers. These estimators are: OLS, Huber's minimax estimator, Wu and Cinar's ARMENSI, and our wavelet-based adaptive robust M-estimator (WARME). Special attention is given to the efficiency of WARME in eliminating the influence of outliers, as well as the adaptiveness to different error distributions.

Twelve error distributions were generated, ranging from light-tailed distributions to heavy-tailed distributions and from symmetric distributions to skewed distributions. They are: (1) Normal distribution N(0, σ^2); (2) Uniform distribution U

$(-a, a)$; (3) t distribution with freedom of degree 2; (4) t distribution with freedom of degree 3; (5) F (5,5) distribution; (6) χ^2 (2) distribution; (7) χ^2 (4) distribution; (8) χ^2 (8) distribution; (9) gamma (2,3) distribution; (10) gamma (2, 5) distribution; (11) Weibull (1, 2) distribution; and (12) Weibull (3, 3) distribution.

All above generated error data are contaminated by 5% isolated outliers generated by the corresponding distribution, whose variance is five times of the original one. The noise-to-signal ratio is set at 0.05. We believe that these distributions and error data can cover most situations occurring in the practical problems.

For the case study, we have assumed a simple isothermal CSTR, in which one reaction $A \Leftrightarrow B$ occurs, and the feed-stream only contains one species. The rate of reaction of disappearance of A is given by

$$r_A = -k[A]^n[B]^m \tag{23}$$

Hence, the rate of appearance of B is

$$r_B = k[A]^n[B]^m \tag{24}$$

The CSTR equations can then be written

$$\frac{dVA}{dt} = F_o A_o - VkA^nB^m - FA$$
$$\frac{dVB}{dt} = VkA^nB^m - FB \tag{25}$$

The values of the variables and parameters for the simulation study are given in Table 1. In order to produce a SISO system, only the inlet concentration of A is assumed to change. Assuming constant volume, the equations can be rewritten as

Table 1. Variable and Parameter Values for Case Study

Parameter	F_0	F	V	A_0	K	n	m
Values	1	1	10	$1 \pm 10\%$	0.134	1.5	-0.5

$$\frac{dA}{dt} = \frac{F_o}{V}A_o - kA^nB^m - \frac{F}{V}A$$
$$\frac{dB}{dt} = kA^nB^m - \frac{F}{V}B \tag{26}$$

If the variation in the inlet concentration of A is assumed to be small, the concentration of B can be approximated to be

$$B = A_o - A \tag{27}$$

As $A_o = 1 \pm 10\%$

$$B \approx 1 - A \tag{28}$$

Hence, the model can be rewritten as

Table 2. Relative Errors and Relative Efficiencies (in Parentheses) of OLS, Huber, ARMENSI and WARME

	OLS	*Huber*	*ARMENSI*	*WARME*
N(0,0.02)	3.5287E−4 (0.5456)	2.4060E−4 (0.8002)	2.1994E−4 (0.8753)	1.9253E−4 (1.0000)
Uniform [−0.02,0.02]	4.7765E−3 (0.7370)	4.5962E−3 (0.7659)	3.5242E−3 (0.9988)	3.5205E−3 (1.0000)
t_2	2.2169E−3 (0.5503)	1.4787E−3 (0.8250)	1.4546E−3 (0.8387)	1.2200E−3 (1.0000)
t_3	3.6081E−1 (0.0073)	2.0552E−2 (0.1288)	1.4960E−2 (0.1769)	2.6470E−3 (1.0000)
F(5,5)	5.5523E−3 (0.7418)	4.2358E−3 (0.97248)	4.2210E−3 (0.97588)	4.1192E−3 (1.0000)
$\chi^2(2)$	2.5811E−2 (0.5341)	1.4342E−2 (0.9613)	1.4035E−2 (0.9823)	1.3787E−2 (1.0000)
$\chi^2(4)$	1.0307E−2 (0.3640)	4.0726E−2 (0.9212)	4.0538E−2 (0.9255)	3.7520E−2 (1.0000)
$\chi^2(8)$	2.2340E−0 (0.0453)	9.5500E−1 (0.1059)	4.2600E−1 (0.2375)	1.0120E−1 (1.0000)
Gamma(2,3)	3.1362E−0 (0.0844)	1.0552E−0 (0.2508)	7.4960E−1 (0.3531)	2.4670E−1 (1.0000)
Gamma(2,5)	2.0950E−1 (0.1303)	1.8660E−1 (0.1463)	3.9690E−2 (0.6878)	2.7300E−2 (1.0000)
Weilbull(1,2)	2.2951E−1 (0.03354)	1.8004E−2 (0.4276)	1.4420E−2 (0.5339)	7.6990E−3 (1.0000)
Weilbull(3,3)	1.0950E−2 (0.09473)	1.1866E−2 (0.08741)	1.1969E−3 (0.8666)	1.0373E−3 (1.0000)

$$\frac{dA}{dt} = \frac{F_o}{V}A_o - kA^n(1-A)^m - \frac{F}{V}A \qquad (29)$$

Using Euler's method for solving ODEs, the equation can be approximated as

$$A_{t+1} = A_t + \left(\frac{F_o}{V}A_o - kA_t^n(1-A_t)^m - \frac{F}{V}A_t\right)\Delta t \quad (30)$$

If the step size $\Delta t = 1$, the equation can be written as

$$A_{t+1} = A_t + \frac{F_o}{V}A_o - kA_t^n(1-A_t)^m - \frac{F}{V}A_t$$

$$A_{t+1} = \left(1-\frac{F}{V}\right)A_t - k(1-A_t)^m A_t^n + \frac{F_o}{V}A_o \qquad (31)$$

A nonlinear dynamic equation can be derived to generate data

$$y(t+1) = ay(t) + by(t)^{1.5}(1-y(t))^{-0.5} + cy(t) + \epsilon(t) \qquad (32)$$

The parameters model values are given as following: $a = 0.9$, $b = -0.134$, $c = 0.1$. The input $u(t)$ is selected between [0.9, 1.3] and $\epsilon(t)$ is certain distribution which is specified as above.

100 samples for each case were generated for each distribution. The empirical relative errors of model parameters vector θ of Expression 32 are calculated by the Euclidean norm

$$\text{error}(t) = \frac{\| \theta - \hat{\theta} \|}{\| \theta \|} \text{or error } (t) = \sum_{i=1}^{m}\left(\frac{\theta_i - \hat{\theta}_i}{\theta_i}\right)^2 \qquad (33)$$

where $m = 3$, $\theta = [a\ b\ c]$, and $\hat{\theta}$ is the estimated value of the model parameter vector.

The relative efficiencies have been obtained by comparing these relative errors to the smallest one of the other estimates for each case studied. The smaller the relative error, the better the model parameter estimation. Also, the larger the relative efficiency, the better the estimator. Results of the four estimates with respect to different error distributions and outliers are reported in Table 2, depicting the empirical relative error norm and relative efficiency, calculated on the basis of 50 Monte-Carlo trails and 100 iterations. Table 3 reports the results on the estimates for some distributions without contaminated outliers.

Table 3. Relative Errors and Relative Efficiencies (in Parentheses) of OLS, Huber, ARMENSI and WARME

	OLS	*Huber*	*ARMENSI*	*WARME*
t_2	8.3096E−3 (0.76272)	6.3380E−3 (0.9999)	6.4410E−3 (0.9840)	6.3379E−3 (1.0000)
Uniform	2.9281E−3 (0.8246)	2.9281E−3 (0.8246)	2.9260E−3 (0.8253)	2.4148E−3 (1.0000)
Normal	2.1876E−3 (1.0000)	2.4667E−3 (0.8868)	3.1082E−3 (0.7248)	2.3245E−3 (0.9411)

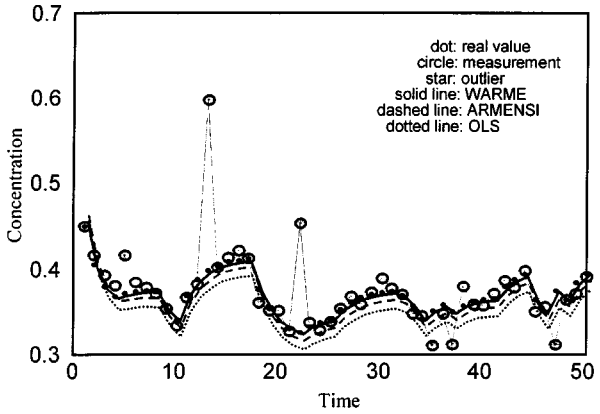

Figure 5. Comparison of different estimation method.

Based on this Monte-Carlo study, for this family of error distributions and outliers, WARME has the best performance. Of course, OLS is the best for normal distribution, due to its characteristics of unbiasness and minimum variance for normal distributions, but it performs very poorly for the case when some outliers occur. The minimax estimator can omit the influence of the outliers, but it only works well for the symmetric distribution, even though it can give acceptable results for cases with no outliers. The ARMENSI performs well for skew distributions, but not as well as WARME does. Figure 5 illustrates the comparative prediction results of OLS, ARMENSI, and WARME methods.

In addition, by using ARMENSI for parameter estimation, one needs to pre-select the order of the polynomials. This means that one needs to have some *a priori* information of the density function of the error (Cobb et al., 1983). Figure 6 illustrates that we can obtain different estimates, by changing the order of the polynomials, in ARMENSI while WARME remains same.

Conclusions

Most industrial and experimental data are contaminated by some anomalous outliers, usually in some unpredicted fashion, which results in the deviation from the assumption of normal error distribution. The performance of conventional system identification methods will deteriorate by using this data for system identification. Robust system identification methods are needed in order to alleviate the effects of the outliers. Several different robust system identification methods have been developed in the past years, mostly in an *ad hoc* manner.

Beginning with a survey of previous work on robust system identification, followed by a discussion of their advantages and disadvantages, a new off-line wavelet-based robust M-estimation procedure (WARME) is developed in this article. The idea is straightforward in the maximum likelihood estimation (MLE) sense, and this approach can be regarded as a generalization of previous robust and nonrobust system identification approaches. The robustness of the proposed approach is determined by its adaptive influence function (IF), which is in turn determined by the residual data themselves

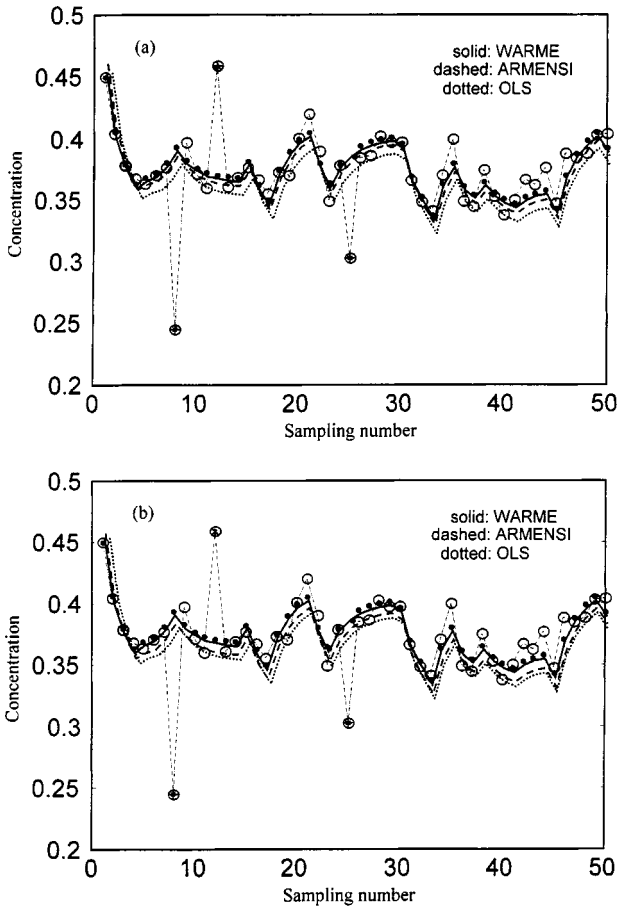

Figure 6. Estimation from different polynomial order, by ARMENSI (a) $n = 30$; (b) $n = 17$.

n: order of polynomial.

rather than an assumption on its density distribution. The wavelet-based density estimation approach is employed to insure an adaptive and more accurate approximation to the p.d.f. of the residuals data. The adaptiveness of the influence function is the striking feature of this approach, while other previous robust approaches are typically based on the assumption on the residual distribution. The performance of this new method has been compared to the conventional ordinary least-squares estimate, as well as other robust procedures including ARMENSI, developed recently in a Monte-Carlo study. This method is a nonparametric approach and is flexible and adaptive to any error distribution without any knowledge of *a priori* information. From a practical viewpoint, it is a more useful and powerful tool for robust nonlinear system identification.

Acknowledgment

The authors gratefully acknowledge the financial support of Australian Research Council (ARC-A2229) and one of the authors (A.A.S.) acknowledges the support by Shiraz University, Iran. We are especially grateful to John Kavanagh for his support and discussion in the case study.

Literature Cited

Andrews, D. F., "A Robust Method for Multiple Linear Regression," *Technometrics*, **16**, 523 (1974).

Albuquerque, J., and L. Biegler, "Data Reconciliation and Gross-Error Detection for Dynamic System," *AIChE J.*, **42**(10), (Oct. 1996).

Butler, R., J. McDonald, R. Nelson, and S. White, "Robust and Partially Adaptive Estimation of Regression Models," *The Rev. Econ. and Stat.*, **72**, 321 (1990).

Cencov, N. N., "Evaluation of an Unknown Distribution Density From Observation," *Sov. Math.* **3**, 1559 (1962).

Cobb, L., P. Koppstein, and N. H. Chen, "Estimation and Moment Recursion Relations for Multimodel Distributions of the Exponential Family," *J. of Amer. Sata. Assn.*, **78**, 381, 124 (Mar. 1983).

Daubechies, I., *Ten Lectures on Wavelets*, Society for International Applied Mathematics (SIAM), Philadelphia (1992).

Forsayth, A. B., "Robust Estimation of Straight Line Regression Coefficients by Minimizing p-th Power Deviations," *Technometrics*, **14**, 159 (1972).

Hall, P., P. Patil, and Y. K. Troung, "On Wavelet Methods for Curve Estimation in Time Series," *Res. Rep. No. SRR*-034-94 *CMA*, Australia National Univ., Canberra, Australia (1994).

Hampel, F. R., E. M. Ronchetti, and W. A. Stahel, *Robust Statistics*: *The Approach Based on Influence Functions*, Wiley, New York (1986).

Hill, R. W., and P. W. Holland, "A Monte Carlo Study of Two Robust Alternatives to Least Squares Regression Estimation," Working Paper 58, *National Bureau of Economic Research*, Cambridge, MA (1974).

Huber, P. J., "Robust Statistics: A Review," *Ann Math. Statist.* **43**, 1041 (1972).

Huber, P. J., *Robust Statistics*, Wiley, New York (1981).

Kerkyacharian, G., and D. Picard, "Density Estimation in Bescov Spaces," *Stat. Probab. Lett.*, **13**, 15 (1992).

Lye, J. N., and V. L. Martin, "Robust Estimation, Nonnormalities, and Generalized Exponential Distributions," *J. of Amer. Sata. Assn.*, **88**, 421, 261 (Mar. 1993).

Mallat, S. G., "A Theory for Multi-Resolution Signal Decomposition: The Wavelet Representation," *IEEE Tran. Anal. Math. Int.*, **PAMI-11**, 674 (1989).

Moberg, T. F., J. S. Ramberg, and R. H. Randles, "An Adaptive Multiple Regression Based on M-Estimators," *Technometrics*, **22** (Mar. 1980).

Narasimhan, S., and R. S. H. Mah, "Generalized Likelihood Ratios for Gross Error Identification in Dynamic Process," *AIChE J.*, **34** 1321 (1988).

Poljak, B. T., and J. A. Z. Tsypkin, "Robust Identification," *Automatica*, **16**, 53 (1980).

Safavi, A. A., "Wavelet Neural Networks and Multiresolution Analysis with Applications to Chemical Engineering," *PhD Thesis*, Univ. of Sydney, Australia (1996).

Safavi, A. A., J. Chen, and J. A. Romagnoli, "Wavelet-Based Density Estimation and Application to Process Monitoring," *AIChE J.*, (May 1997).

Stromberg, A. J., and D. Ruppert, "Breakdown in Nonlinear Regression," *J. of the American Stat. Assn.*, **87**, (Dec. 1992).

Tamhane, A. C., C. Kao, and R. S. H. Mah, "Gross Error Detection in Serially Correlated Process Data 2 Dynamic Systems," *Ind. Eng. Chem. Res.*, **31**, 254 (1992).

Wu, X., and A. Cinar, "An Adaptive Robust M-Estimator for Nonparametric Nonlinear System Identification," *J. Proc. Cont.*, **6**, 233 (1996).

Yuh, L., and R. Hogg, "On Adaptive M-Regression," *Biometrics*, **44**, 433 (June 1988).

Manuscript received Apr. 13, 1999, and revision received Dec. 6, 1999.