Parameter estimation of linear systems with input–output noisy data: A generalized $l_p$ norm approach

Jeng-Ming Chen, Bor-Sen Chen*, Wei-Sheng Chang

Department of Electrical Engineering, National Tsing-Hua University, Hsin Chu, 300, Taiwan ROC

Received 29 October 1992; revised 3 June 1993

Abstract

In this paper, the parameter estimation of linear systems with input–output noisy data is considered. The system input and output are supposed to be corrupted by measurement noises, and the noise distributions are assumed to be unknown. For achieving an efficient parameter estimation, an $l_p$ norm iterative estimation algorithm ($1 < p < \infty$) is proposed. Furthermore, a method for achieving the consistent estimation is presented. Since the exponent $p$ of the $l_p$ norm estimation algorithm is essentially sensitive to the noise distribution, based on the sample kurtosis of the residual, an adequate exponent $p$ can be selected to achieve an efficient parameter estimation at each iteration step. Finally, several simulation results are presented to illustrate the proposed generalized $l_p$ norm iterative estimation algorithm, and we find that the proposed algorithm is a good approach to the system parameter estimation problem with unknown input–output measurement noise distributions.

Zusammenfassung


Résumé

Dans cet article, l’estimation des paramètres de systèmes linéaires avec des données d’entrées–sorties bruitées est étudiée. Les entrées–sorties du système sont supposées être altérées par des bruits de mesure dont les distributions sont supposées inconnues. Pour obtenir une estimation efficace des paramètres, l’utilisation d’un algorithme d’estimation

*Corresponding author.

0165-1684/94/$7.00 © 1994 Elsevier Science B.V. All rights reserved
SSDI 0165-1684(93)00113-Y
1. Introduction

The problem of system parameter estimation is a major field in control and signal processing. In fact, during recent years, many different estimation methods have been proposed. However, in most cases it is assumed that the measurements of the system output are noisy but the measurements of the input to the system are known exactly—see, e.g., [2, 8, 21]. As is well known, in the practical situation, the observed input–output data of an identified system are usually corrupted by measurement noises, and this phenomenon will have an influence on the accuracy of the estimation of system parameters. This paper is concerned with the problem of estimation of system parameters when the input as well as the output measurements are noisy.

The least squares ($l_2$) estimator is generally regarded as the best linear unbiased estimator (BLUE) when the noise distribution is white normal [9]. However, the least squares estimation is far from the optimal in many non-Gaussian situations [7]. Therefore, if the quadratic loss function is not a suitable measure of the loss (i.e., while the noise distribution is not Gaussian), it is necessary to develop an alternative estimation approach.

In fact, for special cases, the least squares estimation ($l_2$), the least absolute value estimation ($l_1$) and the minimax estimation ($l_{\infty}$) methods have been proposed by many authors, e.g., [1, 3, 5, 9]. It is well known that the $l_1$ norm estimators are efficient when the noise follows the Laplace distribution; the $l_2$ norm estimators are efficient when the noise follows the Gaussian distribution; and the $l_{\infty}$ norm estimators are efficient when the noise follows the uniform distribution. Furthermore, it has been shown that the $l_p$ norm estimator is the maximum likelihood estimator when the probability density function of noise is generalized p-Gaussian (gpG) [13]. Therefore, if we want to develop an estimation method which is efficient, a priori information of the noise distribution is necessary.

In general, under the input–output noisy data case, it is difficult to estimate the system parameters. However, only a limited number of estimation methods for systems with input–output noisy data have been proposed, and the quadratic criterion is adopted by most estimation methods [4, 15]. Under the assumption that the true measurement noise distributions are unknown, the purpose of this paper is to study an efficient parameter estimation method using white-noise-corrupted input–output data. Since estimation methods which are efficient for an alternative noise distribution may be quite different, it is desirable to have design procedures which are useful over a range of noise distributions. Therefore, in this paper, a generalized $l_p$ norm method is proposed to solve the system parameter estimation problem with unknown input–output noise distributions. Furthermore, under the input–output noisy data case, the $l_p$ norm estimation algorithm is biased. Hence, in this paper, how to eliminate the estimation bias is also considered.

Because a priori information of the noise distribution is assumed to be unknown and the exponent $p$ of the $l_p$ norm estimation algorithm is essentially sensitive to the noise distribution, it is necessary to search for an adequate value of $p$ in each estimation.
The following assumptions are made:
(A1) \( \{r(t)\} \) and \( \{w(t)\} \) are zero mean white noise sequences with the same statistical distribution, differing only in their variances:
\[
E[r^2(t)] = \sigma_r^2, \quad E[w^2(t)] = \sigma_w^2,
\]
where \( E \) denotes the expectation operator. Furthermore, they are mutually uncorrelated.

(A2) The noise sequences \( \{r(t)\} \) and \( \{w(t)\} \) are uncorrelated with \( \{e(t)\} \).

(A3) The driving input sequence \( \{e(t)\} \) is a stationary ergodic random sequence.

(A4) The structure parameters \( p \) and \( q \) of the identified system are known, and all the zeros of the polynomial \( A(z^{-1}) \) lie strictly inside the unit circle.

If the observation equation (2) is now substituted into (1), then we obtain the following relationship between the variables \( \{u(t)\} \) and \( \{y(t)\} \):
\[
A(z^{-1})y(t) = B(z^{-1})u(t) + v(t),
\]
where
\[
v(t) = A(z^{-1})w(t) - B(z^{-1})r(t).
\]

The problem to be considered is to estimate the system parameters from available data \( \{y(t), u(t), t = 1, \ldots, N\} \).

Define a vector \( v^T(t) \) of noise variables as
\[
v^T(t) = [-w(t-1), \ldots, -w(t-p), r(t), \ldots, r(t-q)],
\]
and we also introduce the following notations:
\[
\phi^T(t) = [-x(t-1), \ldots, -x(t-p), e(t), \ldots, e(t-q)],
\]
\[
\varphi^T(t) = [-y(t-1), \ldots, -y(t-p), u(t), \ldots, u(t-q)].
\]

Then, from Eq. (2), we have
\[
\varphi(t) = \phi(t) + v(t). \quad (4)
\]

Now we define the vector of the unknown true parameters in the system model (1) as
\[
\theta = [a_1, \ldots, a_p, b_0, \ldots, b_q]^T.
\]

Then Eqs. (1) and (3) can be rewritten as
\[
x(t) = \phi^T(t)\theta \quad (5)
\]
and
\[ y(t) = \varphi^T(t)\theta + \xi(t), \tag{6} \]
respectively, where \( \xi(t) = w(t) - y^T(t)\theta \).

Given an equation like (6), the weighted least squares (WLS) estimation of \( \theta \) is obtained by
\[ \hat{\theta}_{WLS} = \left[ \frac{1}{N} \sum_{i=1}^{N} \beta(t) \varphi(t) \varphi^T(t) \right]^{-1} \left[ \frac{1}{N} \sum_{i=1}^{N} \beta(t) \varphi(t) y(t) \right], \tag{7} \]
where \( \beta(t) \) is a weighting function.

It is well known that the estimate (7) is inconsistent since the noise \( \xi(t) \) is nonwhite.

In general, the WLS method is also regarded as the most suitable method of estimating the coefficients in a regression model when the noises are Gaussian distributed. However, in many practical situations the noises are not Gaussian distributed, and in such cases the WLS technique may provide relatively poor estimates of the regression coefficients.

In fact, from the experimental results of Money et al. [16], we know that the \( l_p \) norm method provides alternatives to the least squares (\( l_2 \)) method for estimating the coefficients of a linear regression model; furthermore, the \( l_p \) norm method \( (p \neq 2) \) showed an improvement of efficiency over the least squares method for all non-Gaussian noise distributions. Therefore, for system model (6) with unknown noise distributions, an \( l_p \) norm estimation algorithm will be developed in the next section.

3. Formulation and method of \( l_p \) norm estimation

3.1. Formulation of the \( l_p \) norm estimation problem

Consider the linear system model as (6)
\[ y(t) = \varphi^T(t)\theta + \xi(t). \]

Now the \( l_p \) norm estimation problem is defined as follows:
Find the parameter \( \hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_{p+q+1}]^T \) which minimizes
\[ E_p(\hat{\theta}) = \sum_i |y(t) - \varphi^T(t)\hat{\theta}|^p = \sum_i |\hat{\theta}(t)|^p, \tag{8} \]
where \( \hat{\theta}(t) \) is the residual (i.e., estimation error). Thus, \( \hat{\theta} \) is the \( l_p \) norm estimate of \( \theta \).

The minimization of \( E_p(\hat{\theta}) \) may be achieved by finding the roots of
\[ \frac{\partial E_p(\hat{\theta})}{\partial \hat{\theta}_i} = 0, \quad i = 1, 2, \ldots, p + q + 1, \tag{9} \]
where (9) can be solved iteratively by finding the solution of
\[ \frac{\partial E_p(\hat{\theta})}{\partial \hat{\theta}_i} + \sum_{j=1}^{p+q+1} \frac{\partial^2 E_p(\hat{\theta})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} (\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}) = 0, \tag{10} \]
which is the Newton-Raphson method [18].

Hence, from Eq. (8), we find that (at the \((k+1)\)th iteration)
\[ \frac{\partial E_p(\hat{\theta}^{(k+1)})}{\partial \hat{\theta}_i} = -p \sum_i |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \cdot \varphi_i^T(t) \times \text{sgn} (y(t) - \varphi^T(t)\hat{\theta}^{(k)}) \]
\[ = -p \sum_i W^{(k+1)}(i) |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \varphi_i^T(t), \]
\[ \frac{\partial E_p(\hat{\theta}^{(k+1)})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} = p(p-1) \sum_i |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \cdot \varphi_i^T(t) \varphi_j^T(t) \]
\[ = p(p-1) \sum_i W^{(k+1)}(i) \varphi_i^T(t) \varphi_j^T(t), \tag{11} \]
where
\[ \text{sgn}(n) = \begin{cases} 1, & \text{if } n > 0, \\ -1, & \text{if } n < 0, \\ 0, & \text{if } n = 0, \end{cases} \]
\[ W^{(k+1)}(i) = |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \cdot \varphi_i^T(t), \]
and \( \varphi_i^T(t) \) is the \( i \)th entry of vector \( \varphi^T(t) \).

Substituting (11) into (10), we have
\[ -p \sum_i W^{(k+1)}(i) \varphi_i^T(t) \varphi_j^T(t) \]
\[ + p(p-1) \sum_i W^{(k+1)}(i) \varphi_i^T(t) \varphi_j^T(t) \times (\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}) = 0. \]

Thus we have the following equation:
\[ (p-1) H^{(k+1)}(\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}) = D^{(k+1)} - H^{(k+1)} \hat{\theta}_j^{(k)}, \]
where \( \hat{\theta}(t) \) is the residual (i.e., estimation error). Thus, \( \hat{\theta} \) is the \( l_p \) norm estimate of \( \theta \).

The minimization of \( E_p(\hat{\theta}) \) may be achieved by finding the roots of
\[ \frac{\partial E_p(\hat{\theta})}{\partial \hat{\theta}_i} = 0, \quad i = 1, 2, \ldots, p + q + 1, \tag{9} \]
where (9) can be solved iteratively by finding the solution of
\[ \frac{\partial E_p(\hat{\theta})}{\partial \hat{\theta}_i} + \sum_{j=1}^{p+q+1} \frac{\partial^2 E_p(\hat{\theta})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} (\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}) = 0, \tag{10} \]
which is the Newton-Raphson method [18].

Hence, from Eq. (8), we find that (at the \((k+1)\)th iteration)
\[ \frac{\partial E_p(\hat{\theta}^{(k+1)})}{\partial \hat{\theta}_i} = -p \sum_i |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \cdot \varphi_i^T(t) \times \text{sgn} (y(t) - \varphi^T(t)\hat{\theta}^{(k)}) \]
\[ = -p \sum_i W^{(k+1)}(i) |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^{p-2} \varphi_i^T(t) \varphi_i^T(t), \]
\[ \frac{\partial E_p(\hat{\theta}^{(k+1)})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} = p(p-1) \sum_i |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \cdot \varphi_i^T(t) \varphi_j^T(t) \]
\[ = p(p-1) \sum_i W^{(k+1)}(i) \varphi_i^T(t) \varphi_j^T(t), \tag{11} \]
where
\[ \text{sgn}(n) = \begin{cases} 1, & \text{if } n > 0, \\ -1, & \text{if } n < 0, \\ 0, & \text{if } n = 0, \end{cases} \]
\[ W^{(k+1)}(i) = |y(t) - \varphi^T(t)\hat{\theta}^{(k)}|^p \cdot \varphi_i^T(t), \]
and \( \varphi_i^T(t) \) is the \( i \)th entry of vector \( \varphi^T(t) \).

Substituting (11) into (10), we have
\[ -p \sum_i W^{(k+1)}(i) \varphi_i^T(t) \varphi_j^T(t) \]
\[ + p(p-1) \sum_i W^{(k+1)}(i) \varphi_i^T(t) \varphi_j^T(t) \times (\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}) = 0. \]

Thus we have the following equation:
\[ (p-1) H^{(k+1)}(\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}) = D^{(k+1)} - H^{(k+1)} \hat{\theta}_j^{(k)}, \]
where $H^{(k+1)}$ is a $(p+q+1) \times (p+q+1)$ matrix whose $ij$th entry is $\sum_{t} W^{(k+1)}(t) \phi(t)^{T} \phi(t)$, and

$$D^{(k+1)} = \left[ \sum_{t} W^{(k+1)}(t) y(t) \phi(t)^{T} \right]^{T} \sum_{t} W^{(k+1)}(t) y(t) \phi(t)^{T} + q(t) .$$

It is noted that the matrix $H^{(k+1)}$ is nonsingular. Therefore, we have

$$\hat{\theta}^{(k+1)} = \frac{(p-2) \hat{\theta}^{(k)} + [H^{(k+1)}]^{-1} D^{(k+1)}}{p-1} .$$

Now, if the weighting function $\beta(t)$ of the WLS method (7) is identical with $H^{(k+1)}$, based on Eq. (7), we can obtain the following iteration equation:

$$\hat{\theta}^{(k+1)} = \left[ \sum_{t} W^{(k+1)}(t) \phi(t)^{T} \right]^{-1} \times \left[ \sum_{t} W^{(k+1)}(t) \phi(t) y(t) \right] .$$

(12)

By noting that $H^{(k-1)} = \sum_{t} W^{(k+1)}(t) \phi(t) \phi(t)^{T}$, and $D^{(k+1)} = \sum_{t} W^{(k+1)}(t) \phi(t) y(t)$, from Eq. (12), we have $H^{(k+1)} \hat{\theta}^{(k+1)}_{\text{IRLS}} = D^{(k-1)}$. Hence, we obtain the following equation:

$$\hat{\theta}^{(k+1)} = \frac{(p-2) \hat{\theta}^{(k)} + \hat{\theta}^{(k+1)}_{\text{IRLS}}}{p-1} .$$

(13)

We know that the above algorithm is the Newton type method, and it is a rearranged form of the Newton-Raphson method. Since Newton's method is theoretically the most desirable, in this paper, the parameter estimation algorithm for finding the roots of Eq. (9) is based on the iteration scheme of Eq. (13).

REMARKS. (1) In general, we know that the Newton type method did not always converge for $1 < p \leq \frac{3}{2}$. Thus $p = \frac{3}{2}$ is indeed the critical value for the Newton type method [19]. Therefore, for ensuring convergence, Eq. (13) is used for the case $\frac{3}{2} < p < \infty$.

(2) In fact, the iteration scheme of Eq. (12) is an alternative to the Newton Raphson method for solving the $l_{p}$ norm problem. The $\hat{\theta}^{(k)}_{\text{IRLS}}$ in (12) is called the iteratively reweighted least squares (IRLS) estimate of $\theta$ [20]. In general, the IRLS algorithm converges more slowly than the Newton type method. Therefore, for accelerating the rate of convergence, we adopt the Newton-Raphson method for estimating the system parameters.

Recognizing that, for $1 < p < 2$, $W^{(k+1)}(t) \rightarrow \infty$ as $\hat{\theta}^{(k)}(t) \rightarrow 0$. Thus, in such cases, the IRLS algorithm does not meet its convergence condition – the weights $W^{(k+1)}(t)$ must be bounded for all $\hat{\theta}^{(k)}(t)$ [20].

Therefore, for $1 < p < 2$, $W^{(k+1)}(t)$ can be chosen as follows [20]:

$$W^{(k+1)}(t) = \begin{cases} |\hat{\theta}^{(k)}(t)|^{p-2}, & |\hat{\theta}^{(k)}(t)| \geq \epsilon, \\ \epsilon^{p-2}, & |\hat{\theta}^{(k)}(t)| < \epsilon, \end{cases}$$

(14)

for a small positive number $\epsilon$.

(3) Assume that the sequences $\{\hat{\theta}^{(k)}\}, \{\hat{\theta}^{(k)}_{\text{IRLS}}\}$ converge to $\hat{\theta}^{*}, \hat{\beta}$, respectively ($\frac{3}{2} < p < \infty$); by (13), we have

$$E[\hat{\theta}^{*}] = \frac{(p-2) E[\hat{\theta}^{*}] + E[\hat{\beta}]}{p-1} ,$$

i.e.,

$$E[\hat{\theta}^{*}] = E[\hat{\beta}] .$$

(15)

Since the IRLS method under unknown input-output noise environment is biased, we know that the estimation method (13) is also biased. Therefore, in the next section, a consistent IRLS estimation method will be developed.

3.2. The choice of the exponent $p$

For many years, studies have shown that the 'best' $p$ depends on the kurtosis of the noise distribution. Furthermore, it shows that a longer tailed distribution has a larger kurtosis, i.e., a smaller 'optimal' $p$ [10, 17]. In the paper by Money et al. [10], for $1 < p < \infty$, the authors recommended that an optimal value of $p$ can be expressed as a function of kurtosis $\gamma$, i.e.,

$$p = \frac{9}{\gamma^{2}} + 1 .$$

It should be noted that, for the Gaussian distribution ($\gamma = 3$), this formula suggests the use of $p = 2$.
Since we need to select the value of exponent \( p \) in each iteration step in the proposed \( f_p \) norm estimation algorithm, we indicate the following method of approaching the value of the exponent \( p \) with the help of Money’s formula.

Find the sample kurtosis [17]

\[
\hat{\gamma} = \frac{N \sum_i (v(t) - \bar{v})^4}{\left( \sum_i (v(t) - \bar{v}) \right)^2},
\]

where \( \bar{v} \) is the mean of residual \( v(t) \) and \( N \) is the sample size.

Then we select the exponent \( \hat{p} \) as

\[
\hat{p} = \frac{9}{\hat{\gamma}} + 1. \tag{16}
\]

Now it is interesting to consider the convergence problem of \( \hat{p} \) – whether the value of \( \hat{p} \) converges to a suitable value. For example, does \( \hat{p} \) converge to 2 in the Gaussian case? In fact, this problem was examined by Gonin and Money [6]. Their results showed that the exponent \( \hat{p} \), expressed as a function of the sample kurtosis, is asymptotically normally distributed. Some simulation results of Gonin and Money are given in Table 1. Now, in the following, we will discuss the convergence problem of \( \hat{p} \) for the Gaussian case.

In principle, we know that if \( v(t) \) is a zero mean stationary ergodic random signal, then we have the following results – see, e.g., [16]:

\[
\lim_{N \to \infty} \frac{1}{N \sum_{i=1}^{N} [v(t) - \bar{v}]^2} \stackrel{w.p. 1}{\to} E[v(t)]^2
\]

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} [v(t) - \bar{v}]^4 \stackrel{w.p. 1}{\to} E[v(t)]^4.
\]

Thus we have

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} [v(t) - \bar{v}]^4 \quad \stackrel{w.p. 1}{\to} \quad \frac{E[v(t)]^4}{\left( E[v(t)]^2 \right)^2} = \gamma.
\]

Furthermore, if \( v(t) \) is a Gaussian signal, by [12], we can obtain the following result:

\[
\gamma = \frac{E[v(t)]^4}{\left( E[v(t)]^2 \right)^2} = 3
\]

i.e., we have

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} [v(t) - \bar{v}]^4 \quad \stackrel{w.p. 1}{\to} \quad \gamma = 3.
\]

Therefore, if \( v(t) \) is a Gaussian signal, we obtain the following result:

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} [v(t) - \bar{v}]^4 \quad \stackrel{w.p. 1}{\to} \quad \gamma = 3.
\]

In Section 6, a simulation study will be performed to examine whether \( \hat{p} \) converges to 2 in the Gaussian case.

### 4. Estimation algorithm

Since the parameter estimation of linear systems with input–output noisy data is considered, from Eq. (7), we know that the WLS parameter

<table>
<thead>
<tr>
<th>Sample size ( N )</th>
<th>Normal errors</th>
<th>Laplace errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Formula A</td>
<td>Formula B</td>
</tr>
<tr>
<td>30</td>
<td>2.373</td>
<td>2.242</td>
</tr>
<tr>
<td>50</td>
<td>2.261</td>
<td>2.219</td>
</tr>
<tr>
<td>100</td>
<td>2.117</td>
<td>2.097</td>
</tr>
<tr>
<td>200</td>
<td>2.070</td>
<td>2.056</td>
</tr>
<tr>
<td>400</td>
<td>2.040</td>
<td>2.019</td>
</tr>
</tbody>
</table>

Formula A: \( p = \frac{9}{\hat{\gamma}^2} + 1 \) for \( 1 < p < \infty \).

Formula B: \( p = \frac{6}{\hat{\gamma}} \) for \( 1 < p < 2 \).
estimation method is always inconsistent. Thus, from Eq. (13), it is quite obvious that the Newton-Raphson method is also inconsistent. In order to achieve a consistent parameter estimation, in this section, a consistent WLS estimation is presented.

As is well known, the LS estimate of parameter \( \theta \) of system model (6) is inconsistent. However, based on the assumption of known noise variances, Schneeweiss [15] proposed the following consistent LS estimation method:

\[
\hat{\beta}_{\text{LS}} = \left[ \frac{1}{N} \sum_{t=1}^{N} \varphi(t)\varphi^T(t) - \Sigma \right]^{-1} \left[ \frac{1}{N} \sum_{t=1}^{N} \varphi(t)y(t) \right],
\]

(17)

where the matrix \( \Sigma \) is defined as \( \Sigma = \text{diag} \{ \sigma_{\varphi}^2, \sigma_{\varphi}^2, \ldots \} \). \( \sigma_{\varphi}, \sigma_{\varphi+1} \) are identity matrices of order \( p, q+1 \), respectively.

Note that the statistical property of WLS is similar to LS; thus we propose a consistent WLS estimation method as follows:

\[
\tilde{\beta}_{\text{WLS}} = \left[ \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)\varphi^T(t) - \Lambda \right]^{-1} \times \left[ \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)y(t) \right],
\]

(18)

where \( w(t) \) is the weighting function and the role of the diagonal matrix \( \Lambda \) is similar to the diagonal matrix \( \Sigma \) in Eq. (17). We now briefly analyze the asymptotic property of estimation method \( \tilde{\beta}_{\text{WLS}} \).

Since the estimate \( \tilde{\beta}_{\text{WLS}} \), in (18) is consistent, we have [16]

\[
\lim_{N \to \infty} \left\{ \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)\varphi^T(t) - \Lambda \right\}^{-1} \times \left[ \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)y(t) \right] = \theta,
\]

(19)

with probability one. Moreover, by the assumption of stationary ergodicity, we have [16]

\[
\lim_{N \to \infty} \left\{ \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)\varphi^T(t) \right\} = E[w(t)\varphi(t)\varphi^T(t)] = \mathcal{A}(\theta),
\]

(20)

\[
\lim_{N \to \infty} \left\{ \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)y(t) \right\} = E[w(t)\varphi(t)y(t)],
\]

with probability one. Therefore, based on the asymptotic theory [14], we get

\[
\left\{ E[w(t)\varphi(t)\varphi^T(t)] - \mathcal{A} \right\}^{-1} \left\{ E[w(t)\varphi(t)y(t)] \right\} = \theta,
\]

(21)

i.e.,

\[
E[w(t)\varphi(t)\varphi^T(t)]\theta - \Lambda \theta = E[w(t)\varphi(t)y(t)].
\]

(22)

Now, we rewrite Eq. (22) as follows:

\[
\theta = R^{-1}E[w(t)\varphi(t)y(t)] + R^{-1}\Lambda \theta.
\]

(23)

where \( R = E[w(t)\varphi(t)\varphi^T(t)] \).

Thus, Eq. (23) shows that if we are able to compensate for the estimation bias \( R^{-1}\Lambda \theta \), then a consistent estimate of parameter vector \( \theta \) can be obtained. This is just the principle of bias compensation [15].

Fundamentally, the theoretical values of \( R \), \( \Lambda \) and \( E[w(t)\varphi(t)y(t)] \) cannot be obtained directly, and it should be replaced by sample averaging.

Therefore, based on Eq. (23), we propose the following iteration estimation scheme:

\[
\hat{\theta}_{\text{WLS}}(k+1) = (\hat{R}_{\text{WLS}}^{(k+1)})^{-1} \mathcal{A}^{(k+1)} \hat{\theta}_{\text{WLS}}^{(k)} + \hat{\theta}_{\text{IRLS}}^{(k+1)}.
\]

(24)

where \( \hat{R}_{\text{WLS}}^{(k+1)} = (1/N) \sum_{t=1}^{N} w(t)\varphi(t)\varphi^T(t) \), the value of weighting function \( w(t) \) is defined as (14), and the matrix \( \hat{A}^{(k+1)} \) can be obtained by solving the following equation:

\[
\mathcal{A}^{(k+1)} \hat{\theta}_{\text{WLS}}^{(k+1)} = - \frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)y(t) - \varphi^T(t) \hat{\theta}_{\text{WLS}}^{(k+1)}.
\]

(25)

Thus, based on Eqs. (13) and (24), we propose the following parameter estimation scheme:

\[
\hat{\theta}^{(k+1)} = \frac{(p-2)\hat{\theta}^{(k)} + \hat{\theta}_{\text{WLS}}^{(k+1)}}{p-1}.
\]

(26)

Now, we can develop the following \( l_p \) (1 < p < \infty) norm estimation algorithm, in which \( i \) denotes the iteration step.
\textbf{l}_p \text{ norm algorithm}

Step 1. With the small positive number \( \varepsilon \) and the value of \( p \), set

\[
w_{t+1}^{(i)}(t) = \begin{cases} \hat{\varepsilon}^{(i)}(t) \varepsilon \quad & p \geq 2, \\ \min \{ w_{t+1}^{(i)}(t) \phi(t) \phi(t)^{p-2}, \varepsilon^{p-2} \} & 1 < p < 2, \\ \end{cases}
\]

\( i = 1, 2, \ldots, N. \)

Step 2. Find the matrix \( \hat{A}^{(i+1)} \) via the following equation:

\[
\hat{A}^{(i+1)} \hat{\theta}^{(i+1)} = - \frac{1}{N} \sum_{t=1}^{N} w_{t+1}^{(i)}(t) \phi(t) v(t) - \phi(t) \hat{\theta}^{(i)}.
\]

Step 3. Calculate the WLS estimate \( \hat{\theta}^{(i+1)}_{WLS} \) by the following equation:

\[
\hat{\theta}^{(i+1)}_{WLS} = \left[ \hat{R}^{(i+1)} \right]^{-1} \hat{A}^{(i+1)} \hat{\theta}^{(i)}
\]

\[
+ \left[ \hat{R}^{(i+1)} \right]^{-1} \left[ \frac{1}{N} \sum_{t=1}^{N} w_{t+1}^{(i)}(t) \phi(t) y(t) \right].
\]

Step 4. Calculate the estimate \( \hat{\theta}^{(i+1)} \) as follows:

\[
\hat{\theta}^{(i+1)} = \begin{cases} \hat{\theta}^{(i+1)}_{WLS} & \text{if } 1 < p \leq \frac{3}{2}, \\ \frac{1 - (p - 2) \hat{\theta}^{(i)}_{WLS} + \hat{\theta}^{(i+1)}_{WLS}}{p} & \text{if } \frac{3}{2} < p < \infty . \end{cases}
\]

Therefore, we propose the following main procedure, where \( j \) denotes the iteration step and \( N \) denotes the size of the data set.

\textbf{Main algorithm}

Step 0 (initial estimate). Given a small positive number \( \varepsilon \). Let \( j = 1 \) and \( \hat{\theta}^{(j-1)} \) be the usual least squares estimate of \( \theta \). Use a consistent \( l_2 \) norm estimation method to get parameter estimate \( \hat{\theta}_{l_2} \). Set \( \hat{\theta}^{(j)} = \hat{\theta}_{l_2} \).

Step 1. Calculate the residuals

\[
\hat{\varepsilon}^{(i)}(t) = y(t) - \phi(t) \hat{\theta}^{(i)} , \quad i = 1, \ldots, N.
\]

Step 2. Determine the exponent \( p_{j+1} \) via Eq. (16).

Step 3. Apply the \( l_p \) norm algorithm to find \( \hat{\theta}^{(i+1)}_{l_p} \).

Step 4. Let \( j = j + 1 \). Go to Step 1 until \( \hat{\theta}^{(i+1)}_{l_p} \) appears to have converged.

\textbf{Remark}. Since the \( l_p \) norm estimation algorithm is an iterative scheme for solving the highly nonlinear minimization problem defined by Eq. (8), the convergence of the algorithm to the optimal solution is not always guaranteed. The typical problem of this case is that more than one local minimum may exist. Therefore, the choice of an initial parameter estimate near the optimal solution is crucial to success in finding the global minimum. In this paper, the true distribution of the additive noise is assumed to be unknown and we know that the least squares (\( l_2 \)) estimation method is a popular and easy method; thus, in the initial estimate, we assumed that the noise distribution is Gaussian, i.e., let \( p = 2 \). However, from Eq. (6), we know that \( \xi(t) \) is a correlated random variable; thus, in the initial estimate, the \( l_2 \) solutions are inconsistent. In fact, we know that if the variances of measurement noises are known, it is easy to remove the estimation bias - see, e.g., [4, 15]. Thus, in this paper, the method proposed by Schneeweiss [15] is used to get a better initial estimate.

5. Convergence analysis

In this section, we will analyze the convergent property of the proposed \( l_p \) norm estimation algorithm.

We first state a theorem.

\textbf{THEOREM 1. (Ortega and Rheinboldt [11]). Consider the equation}

\[
X = \Gamma X + B.
\]

where \( \Gamma \) is an \( n \times n \) matrix and \( X, B \) are \( n \times 1 \) vectors. Let \( \{ X^{(i)} \} \) be an \( n \)-dimensional vector sequence obtained from the following sequential algorithm:

\[
X^{(i+1)} = \Gamma X^{(i)} + B , \quad i = 0, 1, 2, \ldots
\]

For any initial vector \( X^{(0)} \) and any vector \( B \), the vector sequence \( \{ X^{(i)} \} \) converges to \( X \) if and only if the following condition is satisfied:

\[
\max_{1 \leq i \leq n} | \lambda_i(\Gamma) | < 1,
\]

where \( \lambda_i(\Gamma) \) denotes an eigenvalue of \( \Gamma \).

Now, the convergence analysis of the proposed \( l_p \) norm estimation algorithm is carried out as follows.
Based on Eq. (23), the iteration estimation scheme is written as
\[
\hat{\theta}^{(i+1)} = R^{-1}A^\dagger\hat{\theta}^{(i)} + R^{-1}E[w(t)\varphi(t)y(t)],
\]
(29)
by Theorem 1, we know that the sequence \(\{\hat{\theta}^{(i+1)}\}\) obtained from (29) converges to 0 if \(\max_{1 \leq i, j \leq n} |\lambda_j (R^{-1}A)| < 1\).

Let \(\lambda\) be an arbitrary eigenvalue of matrix \(R^{-1}A\) and \(e\) the associated eigenvector. Then \(R^{-1}Ae = \lambda e\), i.e., \(Ae = \lambda Re\), and \(e^TR^TRe\). Moreover, from Eqs. (6), (22) and \(R = E[w(t)\varphi(t)\varphi^T(t)]\), we know that the component of matrix \(A\) is the measurement noise, and the component of matrix \(R\) is composed of signal and measurement noise. In general, the power of signal is larger than the power of noise. Thus we obtain the following result:
\[
|\lambda| = \frac{|e^TRe|}{|e^TAe|} < 1,
\]
i.e., the absolute values of eigenvalues of matrix \(R^{-1}A\) are less than unity. By Theorem 1, we can expect the sequence \(\{\hat{\theta}^{(i+1)}\}\) obtained from (29) to converge to \(\theta\).

It is noted that Eq. (24) and Eq. (29) are of similar form. However, in our proposed method (24), both the system parameters and matrix \(R^{-1}A\) are estimated simultaneously. In fact, we know that if the weighting function \(w(t)\) is defined as (14), the IRLS algorithm is convergent. Hence, the major distinction between (24) and (29) is that the matrix \(R^{-1}A\) in Eq. (24) does not remain constant in the iterative process and needs to update continuously.

Note that, for large \(N\), we have
\[
\hat{\theta}^{*} \rightarrow \theta,
\]
\[
\frac{1}{N} \sum_{t=1}^{N} w(t)\varphi(t)y(t) \rightarrow E[w(t)\varphi(t)y(t)].
\]

Since \(\lambda(R^{-1}A)\) is less than unity, it is reasonable that we may expect \(\lambda(R^{-1}A)\) to be less than unity in the iterative process as long as \(\hat{A}\) is not too bad an estimate of \(A\). A reliable approach is to test whether \(\lambda(R^{-1}A)\) is less than unity in each iteration step, although this is usually unnecessary in most circumstances. Moreover, we know that the matrix operator \(R^{-1}A\) with \(|\lambda(R^{-1}A)| < 1\) is referred to as a contraction mapping [11]. Hence the proposed \(l_p\) norm estimation method (24) may be expected to be contractive and can never diverge. As Eq. (25) can only be satisfied with respect to the true parameters, the iterative process (24) will converge to the true parameter \(\theta\).

Furthermore, by [19], we know that the sequence \(\{\hat{\theta}^{(i+1)}\}\) obtained from (26) did not always converge for \(1 < p \leq \frac{3}{2}\). Therefore, for the \(\frac{3}{2} < p < \infty\) case, we can expect that the sequence \(\{\hat{\theta}^{(i+1)}\}\) is convergent. Assume that \(\{\hat{\theta}^{(i+1)}\}\) converges to \(\hat{\theta}^*\). Since the sequence \(\{\hat{\theta}^{(i+1)}\}\) converges to \(\theta\), from Eq. (26), it is expected that the proposed \(l_p\) norm estimation algorithm is a consistent estimator.

REMARK. In general, the proposed \(l_p\) norm algorithm consists of an inner iteration (calculation of \(p\)) and an outer iteration (minimization over the parameters). Since the proposed \(l_p\) norm algorithm is convergent, we know that \(\hat{p}\) is also convergent. Furthermore, since the proposed \(l_p\) norm estimation algorithm is consistent, we know that \(\hat{p}\) will converge to a suitable value, and for the Gaussian case the convergence analysis is given in Section 3.2.

6. Simulation results

In this section, we use the generalized \(l_p\) norm estimation algorithm indicated in the previous section to solve the system parameter estimation problem with unknown input-output measurement noise distributions. Two examples of an ARMA system are illustrated with several different types of noise distribution. The input driving sequence \(\{e(t)\}\) is white Gaussian with zero mean and unit variance. Monte Carlo simulations were performed by generating 50 independent sequences of \(y(t)\), each of length \(N = 500\) samples. We assume that the system input and the system output have the same signal-to-noise ratio (SNR), and we also assume that the measurement noises \(w(t)\) and \(r(t)\) have the same distribution. The results are summarized in the tables to show the true and estimated parameters (mean ± one standard deviation).
Table 2
True and estimated parameters (mean and std.) of type (i) and type (ii) distributions in Example 1

<table>
<thead>
<tr>
<th>True</th>
<th>Estimated (type (ii))</th>
<th>Estimated (type (ii))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std.</td>
</tr>
<tr>
<td>$a_1 = -1.5$</td>
<td>-1.4888</td>
<td>0.0697</td>
</tr>
<tr>
<td>$a_2 = 0.8$</td>
<td>0.7921</td>
<td>0.0656</td>
</tr>
<tr>
<td>$b_0 = 1.0$</td>
<td>1.0203</td>
<td>0.0203</td>
</tr>
<tr>
<td>$b_1 = -1.2$</td>
<td>1.2080</td>
<td>0.0846</td>
</tr>
<tr>
<td>$b_2 = 0.9$</td>
<td>0.9146</td>
<td>0.0527</td>
</tr>
</tbody>
</table>

Table 3
True and estimated parameters (mean and std.) of type (iii) distribution in Example 1

<table>
<thead>
<tr>
<th>True</th>
<th>Estimated (type (iii))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>$a_1 = -1.5$</td>
<td>-1.5152</td>
</tr>
<tr>
<td>$a_2 = 0.8$</td>
<td>0.8114</td>
</tr>
<tr>
<td>$b_0 = 1.0$</td>
<td>1.0186</td>
</tr>
<tr>
<td>$b_1 = -1.2$</td>
<td>-1.2361</td>
</tr>
<tr>
<td>$b_2 = 0.9$</td>
<td>0.9294</td>
</tr>
</tbody>
</table>

**EXAMPLE 1.** The system model is chosen as
\[
x(t) - 1.5x(t-1) + 0.8x(t-2) = e(t) - 1.25e(t-1),
\]
\[
u(t) = e(t) + r(t),
\]
\[
y(t) = x(t) + w(t).
\]

Different types of distribution for measurement noises $w(t)$ and $r(t)$ are given as follows:

(i) zero mean white Gaussian noise with variance 0.01 (SNR = 20 dB);
(ii) white noise uniformly distributed in the interval $[-0.2, 0.2]$ (SNR = 18.75 dB);
(iii) zero mean white contaminated Gaussian noise with distribution of the form $f = \frac{1}{2}N(0, \sigma_1^2) + \frac{1}{2}N(0, 8\sigma_2^2)$ (SNR = 15 dB).

The proposed $l_p$ norm algorithm is used to estimate the system parameters for the three types of noise distribution; the results are summarized in Tables 2 and 3, which show the averaged solution and one standard deviation (std.). For type (i) distribution, the average estimate value of $p$ is 2.014; thus, it has confirmed that the value of $\hat{p}$ converges to 2 in the Gaussian case. The accuracy is chosen to be within $\pm 10^{-6}$. The average numbers of iterations are 8, 9 and 10 for the type (i), (ii) and (iii) distributions, respectively. In general, from the simulation results, we confirm that the proposed algorithm is a good approach to the parameter estimation problem with unknown input–output measurement noise distributions.

**EXAMPLE 2.** The system model is chosen as
\[
x(t) - 1.5x(t-1) + 0.8x(t-2) = e(t) - 1.25e(t-1),
\]
\[
u(t) = e(t) + r(t),
\]
\[
y(t) = x(t) + w(t).
\]

Different types of distribution for measurement noises $w(t)$ and $r(t)$ are given as follows:

(i) zero mean white Gaussian noise with variance 0.0316 (SNR = 15 dB);
(ii) white Gaussian noise with variance 0.1 multiplied by a white noise uniformly distributed in the interval $[-0.5, 0.5]$ (SNR > 10 dB).

The proposed $l_p$ norm algorithm is used to estimate the system parameters for the two types of noise distribution; the results are summarized in Table 4. Similarly, for type (i) distribution, the average estimate value of $p$ is 2.021; thus, it has also confirmed that the value of $\hat{p}$ converges to 2 in the Gaussian case. In this example, the accuracy is chosen to be within $\pm 10^{-5}$ to reduce the numbers of iteration. The average numbers of iterations are 7 and 8 for type (i) and (ii) distributions, respectively. Similarly, from Table 4, it is seen that the system
Table 4
True and estimated parameters (mean and std.) of type (i) and type (ii) distributions in Example 2

<table>
<thead>
<tr>
<th>True</th>
<th>Estimated (type (i))</th>
<th>Estimated (type (ii))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std.</td>
</tr>
<tr>
<td>$a_1 = -1.5$</td>
<td>-1.4566</td>
<td>0.0381</td>
</tr>
<tr>
<td>$a_2 = 0.8$</td>
<td>0.7472</td>
<td>0.0302</td>
</tr>
<tr>
<td>$h_0 = 1.0$</td>
<td>1.0047</td>
<td>0.0187</td>
</tr>
<tr>
<td>$h_1 = -1.25$</td>
<td>-1.2417</td>
<td>0.0438</td>
</tr>
</tbody>
</table>

parameters can be estimated accurately via the proposed estimation algorithm. Thus we confirm that, in the case of unknown input–output measurement noise distributions, the proposed algorithm is a good approach to the parameter estimation problem.

7. Conclusion

In this paper, without knowledge of the distributions of input and output measurement noise, the problem of system parameter estimation has been studied. Since the IRLS algorithm under unknown input–output noise environment is always biased, a technique for eliminating the estimation bias has been presented and a consistent IRLS algorithm has also been developed. Because the Newton–Raphson method generally converges faster than the IRLS method, for $\frac{3}{2} < p < \infty$, a consistent generalized $l_p$ norm estimation algorithm (Newton–Raphson method) has been proposed to estimate the unknown system parameters. Since the true distribution of additive noise is assumed to be unknown and the exponent $p$ is essentially sensitive to the noise distribution, in this paper, how to choose $p$ in the proposed algorithm has also been considered. For several different types of noise distribution, simulation results have illustrated that the unknown parameters can be obtained accurately. Therefore, although a priori information of noise distribution is unknown, the generalized $l_p$ norm estimation algorithm is still a very good approach to the parameter estimation problem with input–output noisy data.

8. Acknowledgments

The authors wish to thank the reviewers and editor for their helpful suggestions and improvements given in the presentation of this paper.

9. References


