Least Squares Identification of Non-Stationary MA Systems

Feng Ding, Yang Shi and Tongwen Chen

Abstract—The correlation analysis based methods are not suitable for identifying parameters of non-stationary MA systems, for which two algorithms are developed, an iterative and a recursive multi-innovation least squares one. The basic idea is to replace unmeasurable noise terms in the information vector by the estimation residuals, which are computed also according to the parameter estimates. This is a hierarchical computation process. Furthermore, the conditions of convergence of the parameter estimation by the recursive algorithm are derived. The simulation results validate the effectiveness of the algorithms proposed.

Keywords: Hierarchical identification, parameter estimation, convergence properties, least squares, AR models, MA models, ARMA models, martingale convergence theorem, multi-innovation identification, auxiliary model identification

I. PROBLEM FORMULATION

The AR, MA and ARMA models are very important in many areas, including signal processing and time series analysis. For decades, a great deal has been published on parameter identification, adaptive filtering and prediction of AR, MA and ARMA models, e.g., [1]–[5], [7], [8]. However, further research is still required for the following reasons: Most contributions assume that the AR, MA and ARMA systems under consideration are stationary and ergodic, i.e., the process noises are stationary and ergodic, which is usually not the case in practice.

Therefore, exploring the estimation algorithms and their properties for the AR, MA and ARMA models with non-stationarity and non-ergodicity is still open and also the goal in this paper. We will frame our study in the identification problems for non-stationary and non-ergodic MA systems, especially the performance analysis of the MA model identification algorithm involved. Note that the methods used can be easily extended to AR and ARMA models.

Consider the following MA (moving average) model:

\[ y(t) = d_0 v(t) + d_1 v(t-1) + \cdots + d_n v(t-n), \quad d_0 = 1, \quad (1) \]

or

\[ y(t) = D(z)v(t), \quad D(z) = d_0 + d_1 z^{-1} + d_2 z^{-2} + \cdots + d_n z^{-n}. \]

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the National Natural Science Foundation of China (60474039).

F. Ding is with the Department of Test and Control Engineering, Nanchang Institute of Aeronautical Technology, Nanchang, P.R. China, and is currently a Research Associate at the University of Alberta, Edmonton, Canada. fding@ece.ualberta.ca

Y. Shi and T. Chen are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4. yangshi@ece.ualberta.ca (Y. Shi), tchen@ece.ualberta.ca (T. Chen).

Here, \( y(t) \) is the system observation data, \( v(t) \) the unmeasurable stochastic noise with zero mean, and \( z^{-1} \) the shift operator \( [z^{-1}v(t) = v(t-1)] \).

In time series analysis, many methods, e.g., correlation analysis based [2]–[4], may be used to estimate the parameters of the MA model in (1) by assuming that \( v(t) \) is stationary and ergodic, i.e.,

\[ \begin{align*}
&C1 \ E[v(t)] = 0; \quad E[v(t)v(j)] = 0, t \neq j; \quad E[v^2(t)] = \sigma^2.
\end{align*} \]

The objective of this paper is to present identification algorithms to estimate the system parameters \( d_i \) of the non-stationary and non-ergodic MA models by using the available observation \( \{y(t)\} \), and to study the properties of the algorithms involved.

Briefly, the paper is organized as follows. Section II discusses the identification problem of stationary MA models based on correlation analysis, and points out that the correlation analysis methods are not suitable for non-stationary and non-ergodic cases. Section III derives an iterative algorithm for identifying MA models. Section IV presents a recursive algorithm for MA models by replacing unmeasurable noise terms in the information vector by the estimation residuals and analyzes its performance. Section V provides an illustrative example to show the effectiveness of the algorithms proposed. Finally, we offer some concluding remarks in Section VI.

II. THE (NON-)STATIONARY MA MODELS

The definitions of system stationarity and ergodicity show that the auto-correlation function \( R(j) := E[y(t)y(t+j)] \) of the time series \( y(t) \) does not depend on \( t \) and equals the time-averaged value, i.e.,

\[ R(j) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} y(t) y(t+j), \quad j = 0, 1, 2, \cdots. \quad (2) \]

This implies that as \( N \to \infty \), the term \( \frac{1}{N} \sum_{t=1}^{N} y(t)y(t+j) \) in the last equation converges to a constant for each \( j \) and that for large \( N, R(j) \) may be computed approximately by

\[ R(j) \approx \frac{1}{N} \sum_{t=1}^{N} y(t)y(t+j). \quad (3) \]

In order to show that the correlation analysis based methods are not suitable for identifying the parameters of non-stationary MA models, let us begin with a 2-order MA model, i.e., \( n = 2 \). From (1), we have

\[ \begin{align*}
&y(t) + y(t-j) = d_0 v(t-j) + d_1 v(t-j) v(t-j-1) + d_2 v(t-j) v(t-j-2) \quad &+ d_0 d_1 v(t-j) v(t-j-1) v(t) \quad &+ d_2 v(t-j) v(t-j-1) v(t-j) + d_1 d_2 v(t-j) v(t-j-1) v(t-j-2) \quad &+ d_1 d_0 v(t-j) v(t-j-1) v(t-j-2) v(t) \quad &+ d_2^2 v(t-j) v(t-j-2) v(t),
\end{align*} \]

where \( d_0, d_1, d_2 \) are the parameters of the MA model.
Taking the expectation on both sides of the above equation and using the stationary assumption in (C1), when $j = 0, 1, 2$, we obtain $n + 1 = 3$ equations:

$$\begin{align*}
\begin{cases}
    d_0^2 \sigma^2 + d_1^2 \sigma^2 + d_2^2 \sigma^2 = R(0), \\
    d_0 d_1 \sigma^2 + d_1 d_2 \sigma^2 = R(1), \\
    d_0 d_2 \sigma^2 = R(2).
\end{cases}
\end{align*}$$

Then the $n + 1 = 3$ unknown parameters $d_1, d_2$ and $\sigma^2$ can be solved from these $n + 1 = 3$ equations since the correlation functions $R(j), j = 0, 1, \cdots, n$, are available in terms of (3). However, if $\{v(t)\}$ is a non-stationary and uncorrelated stochastic noise sequence with zero mean and time-varying variance $\sigma^2(t)$, i.e.,

$$\begin{align*}
(C2) \ E[v(t)] = 0; \ E[v(t)v(j)] = 0, j \neq t; \ E[v^2(t)] = \sigma^2(t),
\end{align*}$$

then, the correlation functions of $y(t)$ also depend on $t$, denoted by $R(t, j) := E[y(t)y(t + j)]$. A similar derivation of (4) yields

$$\begin{align*}
\begin{cases}
    d_0^2 \sigma^2 + d_1^2 \sigma^2(t - 1) + d_2^2 \sigma^2(t - 2) = R(t, 0), \\
    d_0 d_1 \sigma^2(t - 1) + d_1 d_2 \sigma^2(t - 2) = R(t, 1, 1), \\
    d_0 d_2 \sigma^2(t - 2) = R(t - 2, 2).
\end{cases}
\end{align*}$$

These $n + 1 = 3$ equations contain $3n + 2 = 8$ unknowns: $d_1, d_2, \sigma^2(t), \sigma^2(t - 1), \sigma^2(t - 2)$ and $R(t, 0), R(t, 1, 1)$ and $R(t - 2, 2)$ since the noise variances are unknown and the correlation functions $R(t, i)$ of the non-stationary process cannot be obtained from (2). Even if $R(t, i)$ are available by using some other ways, it is impossible to solve the parameter estimates $d_1$ and $d_2$ from (5) because (5) still has $2n + 1 = 5$ unknowns. Therefore, the correlation-analysis methods are not suitable for identifying non-stationary MA and ARMA processes. Next, we will discuss identification algorithms of non-stationary MA processes.

### III. The Iterative Algorithm

Define the parameter vector $\theta$ and information vector $\varphi_0(t)$ as

$$\begin{align*}
\theta &= [d_1, d_2, \cdots, d_n]^T \in \mathbb{R}^n, \\
\varphi_0(t) &= [v(t - 1), v(t - 2), \cdots, v(t - n)]^T \in \mathbb{R}^n,
\end{align*}$$

and

$$\begin{align*}
Y(t) &= [y(t), y(t - 1), \cdots, y(t - p + 1)]^T \in \mathbb{R}^p, \\
V(t) &= [v(t), v(t - 1), \cdots, v(t - p + 1)]^T \in \mathbb{R}^p, \\
\Gamma_0(t) &= [V(t - 1), V(t - 2), \cdots, V(t - n)] \\
&= [\varphi_0^T(t), \varphi_0^T(t - 1), \cdots, \varphi_0^T(t - p + 1)]^T.
\end{align*}$$

From (1), we easily get

$$\begin{align*}
y(t) &= \varphi_0^T(t) \theta + v(t), \\
Y(t) &= \Gamma_0(t) \theta + V(t),
\end{align*}$$

where the superscript $T$ denotes the matrix transpose; $\theta$ is the parameter vector to be identified, and $p$ may be known as the data length ($t \geq p \gg n$). Form a quadratic criterion function

$$J(\theta) = \|Y(t) - \Gamma_0(t) \theta\|^2,$$

where $\|X\|^2 = \text{tr}[XX^T]$ represents the norm of the matrix $X$. Minimizing $J(\theta)$ gives the least-squares estimate:

$$\hat{\theta} = [\Gamma_0^T(t) \Gamma_0(t)]^{-1} \Gamma_0^T(t) Y(t).$$

However, a difficulty arises because $V(t - i)$ in $\Gamma_0(t)$ is unavailable; so it is impossible to compute the estimate $\hat{\theta}$ by (8). Our approach is based on the iterative (hierarchical) identification principle: Let $k = 1, 2, 3, \cdots$, and $\hat{\theta}_k$ denote the estimate of $\theta$ at iteration $k$, then the unknown variable $V(t - i)$ can be computed/estimated by

$$\hat{V}_k(t - i) = Y(t - i) - \hat{\Gamma}_{k-1}(t - i) \hat{\theta}_{k-1}, \; i = 0, 1, \cdots, n$$

with

$$\hat{\Gamma}_k(t) := [\hat{V}_k(t - 1), \hat{V}_k(t - 2), \cdots, \hat{V}_k(t - n)] \in \mathbb{R}^{p \times n}.$$ 

Based on (8) and replacing $\Gamma_0(t)$ by $\hat{\Gamma}_k(t)$, the iterative solution $\hat{\theta}_k$ of $\theta$ may be also be computed by

$$\hat{\theta}_k = [\hat{\Gamma}_k^T(t) \hat{\Gamma}_k(t)]^{-1} \hat{\Gamma}_k^T(t) Y(t), \; k = 1, 2, 3, \cdots.$$ (11)

Equations (9)-(11) are referred to as the least-squares iterative (or iterative least-squares) identification algorithm for MA systems, MA-LSI algorithm for short.

The MA-LSI algorithm employs the idea of updating the estimate $\hat{\theta}$ using a fixed data batch with a finite length $p$. In this paper, in order to distinguish on-line from off-line calculation, we use iterative with subscript $k$, e.g., $\hat{\theta}_k$, for off-line algorithms, and recursive with no subscript, e.g., $\hat{\theta}(t)$ to be given later, for on-line ones. We imply that a recursive algorithm can be on-line implemented, but an iterative one cannot. For a recursive algorithm, new information (input and/or output data) is always used to recursively computes the parameter estimates at each step as time increases.

To initialize the algorithm in (9) to (11), we take $\hat{\Gamma}_0(t) = 0$ and $\hat{\theta}_0 = 10^{-6} \mathbf{1}_n$ with $\mathbf{1}_n$ being an $n$-dimensional column vector whose elements are 1.

To summarize, we list the steps involved in the MA-LSI algorithm to compute $\hat{\theta}_k$ as $k$ increases:

1. Collect the observation data $\{y(t)\}$, select data length $p$, and form $Y(t)$ by (6).
2. To initialize, let $k = 1$ and $\hat{\theta}_0 = 10^{-6} \mathbf{1}_n$.
3. Compute $\hat{V}_k(t)$ by (9), form $\hat{\Gamma}_k(t)$ by (10).
4. Compute the estimate $\hat{\theta}_k$ by (11).
5. Compare $\hat{\theta}_k$ with $\hat{\theta}_{k-1}$; if they are sufficiently close, or for some pre-set small $\varepsilon$, if

$$\|\hat{\theta}_k - \hat{\theta}_{k-1}\|^2 \leq \varepsilon,$$

then terminate the procedure and obtain the iterative times $k$ and estimate $\hat{\theta}_k$; otherwise, increment $k$ by 1 and go to step 3.

This MA-LSI iterative algorithm cannot guarantee that $\hat{\theta}_k$ converges to $\theta$ in that it uses the finite data set. Next, we derive a recursive algorithm of estimating $\theta$. 

4779
IV. THE RECURSIVE ALGORITHM

Let us introduce some notation first. \(|X| = \det[X]\) represents the determinant of the matrix \(X\); \(\lambda_{\text{max}}[X]\) and \(\lambda_{\text{min}}[X]\) represent the maximum and minimum eigenvalues of \(X\), respectively; \(f(t) = o(g(t))\) represents \(f(t)/g(t) \to 0\) as \(t \to \infty\); for \(g(t) \geq 0\), we write \(f(t) = O(g(t))\) or \(f(t) \sim g(t)\) if there exists a positive constant \(\delta_i\) such that \(|f(t)| \leq \delta_i g(t)\).

As pointed out in the preceding section, the MA-LSI algorithm uses batch data identification and thus is not suitable for on-line identification. Moreover, the major drawback is that it requires computing matrix inversion at each step. In this section, we derive a recursive identification algorithm that can be on-line implemented.

As in the iterative algorithm, the unknown \(V(t-i)\) in the matrix \(\Gamma_0(t)\) are replaced by their estimates \(\hat{V}(t-i)\). Let \(\hat{\theta}(t)\) denote the estimate of \(\theta\) at time \(t\), and use \(\Gamma(t)\) as \(\Gamma_0(t)\), then according to the least squares principle [6], it is not difficult to get the following recursive least squares algorithm of estimating \(\theta\) based on the noise estimation,

\[
\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)\Gamma^T(t) \cdot [Y(t) - \Gamma(t)\hat{\theta}(t-1)],
\]

\[
P^{-1}(t) = P^{-1}(t-1) + \Gamma^T(t)\Gamma(t),
\]

\[
\hat{V}(t-i) = Y(t-i) - \Gamma(t-i)\hat{\theta}(t),
\]

\[
\Gamma(t) = [\hat{V}(t-1), \hat{V}(t-2), \ldots, \hat{V}(t-n)]^T
\]

\[
= : \{\phi(t), \phi(t-1), \ldots, \phi(t-p+1)\}^T.
\]

Here, we take \(P(0) = \rho_0 I\) with \(\rho_0\) being a large positive number, e.g., \(\rho_0 = 10^9\), \(\hat{\theta}(0) = 1_n/\rho_0\), and we refer to \(\hat{V}(t-i)\) as the estimate of \(V(t-i)\).

As \(p = 1\), \(e(t) := y(t) - \phi^T(t)\hat{\theta}(t-1) \in \mathbb{R}^1\) is called the innovation [6], and \(E(t) := Y(t) - \Phi^T(t)\hat{\theta}(t-1) \in \mathbb{R}^p\) may be referred to as the innovation vector, i.e., multi-innovation. \(p\) here may be also known as the innovation length, and (12)-(15) is also called the multi-innovation least squares algorithm of identifying MA models, the MA-MLS algorithm for short.

This MA-MLS algorithm performs a hierarchical recursive computation procedure because the noise estimates \(\hat{V}(t-i)\) rely on the parameter estimates \(\hat{\theta}(t)\), see Equation (14), and the parameter estimates \(\hat{\theta}(t)\) also rely on the noise estimates \(\hat{V}(t-i)\), see Equations (15) and (12).

The algorithm in (12)-(15) is simple and easy to implement on-line. However, a very important question posted here is: Under what conditions can the parameter estimation vector \(\hat{\theta}(t)\) converge to the true parameter vector \(\theta\)? The following answers this question.

Define

\[
P_0^{-1}(t) := P_0^{-1}(t-1) + \Gamma_0^T(t)\Gamma_0(t), \quad P_0(0) = \rho_0 I; \quad r_0(t) := \text{tr}[P_0^{-1}(t)], \quad r(t) := \text{tr}[P^{-1}(t)].
\]

It follows that

\[
[P^{-1}(t)] \leq r(t), \quad r(t) \geq [P^{-1}(t)]^{1/n};
\]

\[
r(t) \leq n\lambda_{\text{max}}[P^{-1}(t)]; \quad \ln[P^{-1}(t)] = O(\ln r(t)).
\]
where

\[ S(t) := 2 \sum_{i=1}^{j} \theta^T(i) \tilde{Y}(i), \quad \text{a.s.,} \]

\[ \tilde{Y}(t) := \frac{1}{2} \Gamma(t) \tilde{\theta}(t) + [Y(t) - \Gamma(t) \tilde{\theta}(t) - V(t)], \quad (19) \]

\[ \tilde{U}(t) := -\Gamma(t) \tilde{\theta}(t). \quad (20) \]

Here, (A3) guarantees that \( S(t) \geq 0 \). It is obvious that \( v(t) \) is non-stationary and non-ergodic.

**Proof** Define the innovation vector \( E(t) \) and residue vector \( \tilde{V}(t) \) as follows:

\[ E(t) := Y(t) - \Gamma(t) \tilde{\theta}(t-1), \quad (21) \]

\[ \tilde{V}(t) := Y(t) - \Gamma(t) \tilde{\theta}(t), \quad (22) \]

It follows that

\[ \tilde{V}(t) = \left[ I - \Gamma(t)p(t)\Gamma^T(t) \right] E(t) = \left[ I + \Gamma(t)p(t)\Gamma^T(t) \right]^{-1} E(t). \quad (23) \]

Substituting (12) into (17) and using (21)-(23), it is not difficult to get

\[ \tilde{\theta}(t) = \tilde{\theta}(t-1) + P(t)\Gamma^T(t)E(t) = \tilde{\theta}(t-1) + P(t-1)\Gamma^T(t)\tilde{V}(t), \quad (24) \]

or

\[ P^{-1}(t-1) \tilde{\theta}(t) = \tilde{\theta}(t-1) + P(t-1)\Gamma^T(t)\tilde{V}(t). \quad (25) \]

Pre-multiplying (25) by \( \tilde{\theta}^T(t) \) and using (24) yield

\[ \tilde{\theta}^T(t)P^{-1}(t-1) \tilde{\theta}(t) = \tilde{\theta}^T(t-1) + P(t-1)\Gamma^T(t)\tilde{V}(t)^T. \]

Using (13), we have

\[ \tilde{\theta}^T(t)P^{-1}(t-1) \tilde{\theta}(t) = \tilde{\theta}^T(t-1) + \tilde{\theta}^T(t-1) + \tilde{\theta}^T(t)(\Gamma^T(t)\tilde{V}(t)). \]

Using (13), (21) to (24), from (18), we have

\[ W(t) = W(t-1) + \tilde{\theta}^T(t)\Gamma^T(t)\Gamma(t)\tilde{\theta}(t) + \tilde{\theta}^T(t)\Gamma(t)\tilde{\theta}(t-1) + \tilde{\theta}^T(t)(\Gamma(t)\tilde{V}(t)). \]

Using (19), (20), (23) and (24), we have

\[ W(t) \leq W(t-1) - 2\tilde{\theta}^T(t)\tilde{Y}(t) + 2\tilde{\theta}^T(t-1) + P(t)\Gamma^T(t)E(t) \]

\[ \leq W(t-1) - 2\tilde{\theta}^T(t)\tilde{Y}(t) + 2\tilde{\theta}^T(t-1) + P(t)\Gamma^T(t)\tilde{V}(t). \]

(26)

Since \( \Gamma(t) \tilde{\theta}(t-1) \), \( E(t) \), \( \Gamma(t)p(t)\Gamma^T(t) \) are \( \mathcal{F}_{t-1} \)-measurable, taking the conditional expectation of both sides of (26) with respect to \( \mathcal{F}_{t-1} \) and using (A1)-(A2) give

\[ E[\tilde{W}(t)\mid \mathcal{F}_{t-1}] \leq W(t-1) - 2E[\tilde{\theta}^T(t)\tilde{Y}(t)\mid \mathcal{F}_{t-1}] \]

\[ + 2p \sum_{i=0}^{p-1} \tilde{\phi}^T(i)p(t)\tilde{\phi}(t-i)\frac{\partial^2}{\partial \tilde{\phi}(t-i)^2}, \quad \text{a.s.} \quad (27) \]

Since

\[ D(z)[\tilde{V}(t) - V(t)] = -\Gamma(t) \tilde{\theta}(t) = \tilde{U}(t), \quad (28) \]

(20), (28) and (22), from (19), we get

\[ \tilde{V}(t) = \frac{1}{2} \Gamma(t) \tilde{\theta}(t) + [\tilde{V}(t) - V(t)] \]

\[ \leq \left[ \frac{1}{D(z)} - \frac{1}{2} \right] \tilde{U}(t) + \frac{\rho}{2} \tilde{U}(t) \]

\[ =: \tilde{Y}_1(t) + \frac{\rho}{2} \tilde{U}(t), \]

where

\[ \tilde{Y}_1(t) = H_1(z) \tilde{U}(t), \quad H_1(z) = \frac{1}{D(z)} - \frac{1}{2}. \]

Here \( \tilde{Y}(t) \) may be regarded as the output of the transfer function \( H_1(z) \) driven by \( \tilde{U}(t) \). Since \( H(z) \) is a strictly positive real function, there exists a constant \( \rho > 0 \) such that \( H_1(z) \) is also strictly positive real. Referring to Appendix C in [3], we can draw that the following inequalities hold

\[ \sum_{i=1}^{T} \tilde{U}^T(i)\tilde{Y}_1(i) \geq 0, \quad \text{a.s.} \]

\[ S(t) = 2 \sum_{i=1}^{T} \tilde{U}^T(i)\tilde{Y}_1(i) + \rho \sum_{i=1}^{T} \tilde{U}^2(i) \geq 0, \quad \text{a.s.} \quad (29) \]

Adding both sides of (27) by \( S(t) \) gives the conclusion of Lemma 2.

Next, we shall prove the main results of this paper by constituting a martingale process and by using stochastic process theory and the martingale convergence theorem (Lemma D.5.3 in [3]).

**Theorem 1:** For the system in (6) or (7), assume that (A1)-(A3) hold, and \( D(z) \) is stable, i.e., all zeros of \( D(z) \) are inside the unit circle. Then for any \( \beta > 1 \), the parameter estimation error by the MA-MILS algorithm in (12)-(15) satisfies:

\[ \| \tilde{\theta}(t) - \theta \|^2 = O \left( \frac{[\ln r_0(t)]^\beta}{\lambda_{\min}[P_0^{-1}(t)]} \right), \quad \text{a.s.} \]
Proof From the definition of \( W(t) \), we have
\[
\|\hat{\theta}(t)\|^2 \leq \frac{\hat{\theta}^T(t) P^{-1}(t) \hat{\theta}(t)}{\lambda_{\min}(P^{-1}(t))} = \frac{W(t)}{\lambda_{\min}(P^{-1}(t))}.
\] (30)
Let
\[
Z(t) = \frac{W(t) + S(t)}{\ln|\lambda|^\beta}.
\]
Since \( \ln r(t) \) is non-decreasing, according to Lemma 2, we have
\[
E[Z(t)|\mathcal{F}_{t-1}] \leq \frac{W(t-1) + S(t-1)}{\ln|\lambda|^\beta} + 2p \sum_{i=0}^{p-1} \frac{\phi^T(t-i)P(t)\phi(t-i)}{\ln|\lambda|^\beta} \sigma_i^2
\]
\[
\leq Z(t-1) + 2p \sum_{i=0}^{p-1} \frac{\phi^T(t-i)P(t)\phi(t-i)}{\ln|\lambda|^\beta} \sigma_i^2. \tag{31}
\]
Using Lemma 1, it is clear that the sum for \( t \) from 1 to \( \infty \) of the last term on the right-hand side of (31) is finite. Now applying the Martingale convergence theorem (Lemma D.5.3 in [3]) to (31), we conclude that \( Z(t) \) converges a.s. to a finite random variable, say, \( Z_0 \); i.e.,
\[
Z(t) = \frac{W(t) + S(t)}{\ln|\lambda|^\beta} \to Z_0 < \infty, \text{ a.s.}
\]
or
\[
W(t) = O(\ln|\lambda|^\beta), \text{ a.s.}, \quad S(t) = O(\ln|\lambda|^\beta), \text{ a.s.} \tag{32}
\]
Since \( H(z) \) is a strictly positive real function, from (29), it follows that
\[
\sum_{i=1}^{t} \|\tilde{U}(i)\|^2 = \frac{O(\ln|\lambda|^\beta)}. \tag{33}
\]
From (30), (32), we have
\[
\|\hat{\theta}(t)\|^2 = O\left(\frac{\ln|\lambda|^\beta}{\lambda_{\min}(P^{-1}(t))}\right), \text{ a.s.}, \text{ for any } \beta > 1.
\]
Since \( D(z) \) is stable, according to Lemma B.3.3 in [3] and (28), there exist positive constants \( k_1 \) and \( k_2 \) such that
\[
\sum_{i=1}^{t} \|\tilde{V}(i) - V(i)\|^2 \leq k_1 \sum_{i=1}^{t} \|\tilde{U}(i)\|^2 + k_2
\]
\[
= O(\ln|\lambda|^\beta). \tag{34}
\]
Now we prove \( r(t) = O(\lambda_{\min}(P^{-1}(t))) = O(\lambda_{\min}(P_0^{-1}(t))) \). Define the vector error \( \Gamma(t) \) as follows [refer to the definitions of \( \Gamma \) and \( \Gamma_0 \)]:
\[
\Gamma(t) := \Gamma(t) - \Gamma_0(t)
\]
\[
= \tilde{V}(t-1) - V(t-1), \ldots, \tilde{V}(t-n) - V(t-n).
\]
Hence, for any \( \beta > 1 \), we have
\[
\sum_{i=1}^{t} \|\tilde{\Gamma}(i)\|^2 = \sum_{i=1}^{t} \sum_{j=1}^{n} \|\tilde{V}(i-j) - V(i-j)\|^2
\]
\[
= O(\ln|\lambda|^\beta),
\]
\[
r(t) \leq 2r_0(t) + 2 \sum_{i=1}^{t} \|\tilde{\Gamma}(i)\|^2
\]
\[
= 2r_0(t) + O(\ln|\lambda|^\beta) = O(r_0(t)), \text{ a.s.}
\]
For any vector \( \omega \in \mathbb{R}^n \) with \( \|\omega\| = 1 \), we have
\[
\sum_{i=1}^{t} \|\Gamma(i)\omega\|^2 \leq \sum_{i=1}^{t} \|\Gamma_0(i)\omega - \tilde{\Gamma}(i)\omega\|^2
\]
\[
\leq 2 \sum_{i=1}^{t} \|\Gamma_0(i)\omega\|^2 + 2 \sum_{i=1}^{t} \|\tilde{\Gamma}(i)\|^2
\]
\[
= 2 \sum_{i=1}^{t} \|\Gamma_0(i)\omega\|^2 + O(\ln|\lambda|^\beta).
\]
Thus
\[
\lambda_{\min}(P^{-1}(t)) \leq 2\lambda_{\min}(P_0^{-1}(t)) + O(\lambda_{\min}(P_0^{-1}(t)))
\]
\[
= O(\lambda_{\min}(P_0^{-1}(t))).
\]
Hence, it is not difficult to get
\[
\|\hat{\theta}(t) - \theta\|^2 = O\left(\frac{\ln|\lambda|^\beta}{\lambda_{\min}(P_0^{-1}(t))}\right), \text{ a.s., for any } \beta > 1.
\]
This proves Theorem 1. \( \square \)

Moreover, we may draw the following corollary from Theorem 1.

Assume that there exist positive constants \( c_0, c_1, c_2 \) and \( t_0 \) such that for any \( t \geq t_0 \), the following generalized persistent excitation condition (unbounded condition number) holds:
\[
(C1) \quad c_1 I \leq \frac{1}{2} \sum_{i=1}^{t} \Gamma_0(i)^T\Gamma_0(i) \leq c_2 t_0^\gamma I, \text{ a.s.}
\]
Then for any \( \beta > 1 \), we have
\[
\|\hat{\theta}(t) - \theta\|^2 = O\left(\frac{\ln|\lambda|^\beta}{t}\right) \to 0, \text{ a.s.}
\]
Since \( \ln t = o(t^\gamma) \) (e.g. arbitrary small), \( \|\hat{\theta}(t) - \theta\| \) converges to zero approximately at the rate of \( 1/\sqrt{t^{1-\gamma}} \) for non-stationary noise.

Taking \( p = 1 \) in the MA-MILS algorithm, we get a simple recursive least squares algorithm based on the noise estimation, the MA-RLS algorithm for short,
\[
\hat{\theta}(t) = \hat{\theta}(t-1) + P(t) \phi(t) [y(t) - \phi^T(t) \hat{\theta}(t-1)],
\]
\[
P(t) = P(t-1) - P(t-1) \phi(t) \phi^T(t) P(t-1) \frac{1}{1 + \phi^T(t) P(t-1) \phi(t)},
\]
\[
\hat{\phi}(t) = y(t) - \phi^T(t) \hat{\theta}(t),
\]
\[
\phi(t) = [\hat{\phi}(t-1), \hat{\phi}(t-2), \ldots, \hat{\phi}(t-n)]^T \in \mathbb{R}^n.
\]

V. EXAMPLE

An example is given to demonstrate the effectiveness of the proposed algorithm. Consider the following simulation plant:
\[
y(t) = D(z) v(t),
\]
\[
D(z) = 1 + d_1 z^{-1} + d_2 z^{-2}
\]
\[
= 1 + 0.412 z^{-1} + 0.309 z^{-2},
\]
\[
\theta = [d_1, d_2]^T = [0.412, 0.309]^T.
\]
\{v(t)\} is taken as a white noise sequence with zero mean and time-varying variance. Apply the MA-MILS algorithm with \( p = 1 \) to estimate the parameters of this MA model, the parameter estimates \( d_i \) and their errors are shown in Table I, and the parameter estimation error \( \delta \) versus \( t \) is shown in Fig. 1, where \( \delta = \| \hat{\theta}(t) - \theta \| / \| \theta \| \) is the relative parameter estimation errors.

<table>
<thead>
<tr>
<th>( t )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( \delta (%) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.54958</td>
<td>0.53833</td>
<td>51.92873</td>
</tr>
<tr>
<td>200</td>
<td>0.52710</td>
<td>0.43024</td>
<td>32.46150</td>
</tr>
<tr>
<td>300</td>
<td>0.44045</td>
<td>0.36930</td>
<td>12.94705</td>
</tr>
<tr>
<td>500</td>
<td>0.41298</td>
<td>0.35975</td>
<td>9.85596</td>
</tr>
<tr>
<td>800</td>
<td>0.42472</td>
<td>0.36475</td>
<td>11.0289</td>
</tr>
<tr>
<td>1000</td>
<td>0.43643</td>
<td>0.37015</td>
<td>12.7860</td>
</tr>
<tr>
<td>1500</td>
<td>0.4297</td>
<td>0.35508</td>
<td>10.22612</td>
</tr>
<tr>
<td>2000</td>
<td>0.41717</td>
<td>0.36141</td>
<td>7.93853</td>
</tr>
<tr>
<td>2500</td>
<td>0.41748</td>
<td>0.35508</td>
<td>7.12897</td>
</tr>
<tr>
<td>3000</td>
<td>0.40916</td>
<td>0.34563</td>
<td>5.8897</td>
</tr>
<tr>
<td>3500</td>
<td>0.40916</td>
<td>0.34563</td>
<td>5.8897</td>
</tr>
<tr>
<td>4000</td>
<td>0.41801</td>
<td>0.32058</td>
<td>4.69745</td>
</tr>
<tr>
<td>4500</td>
<td>0.41634</td>
<td>0.31570</td>
<td>3.8497</td>
</tr>
<tr>
<td>5000</td>
<td>0.41741</td>
<td>0.31518</td>
<td>3.4559</td>
</tr>
<tr>
<td>5500</td>
<td>0.42543</td>
<td>0.30820</td>
<td>2.6128</td>
</tr>
<tr>
<td>6000</td>
<td>0.41706</td>
<td>0.31313</td>
<td>1.0043</td>
</tr>
<tr>
<td>6500</td>
<td>0.41579</td>
<td>0.31373</td>
<td>1.1846</td>
</tr>
<tr>
<td>7000</td>
<td>0.41900</td>
<td>0.31742</td>
<td>1.7285</td>
</tr>
<tr>
<td>7500</td>
<td>0.41181</td>
<td>0.31066</td>
<td>0.3246</td>
</tr>
<tr>
<td>8000</td>
<td>0.40790</td>
<td>0.31275</td>
<td>1.0737</td>
</tr>
<tr>
<td>8500</td>
<td>0.41066</td>
<td>0.31346</td>
<td>0.9445</td>
</tr>
<tr>
<td>9000</td>
<td>0.41200</td>
<td>0.31338</td>
<td>0.8497</td>
</tr>
<tr>
<td>9500</td>
<td>0.41068</td>
<td>0.31055</td>
<td>0.3956</td>
</tr>
<tr>
<td>10000</td>
<td>0.40994</td>
<td>0.30678</td>
<td>0.58897</td>
</tr>
</tbody>
</table>

**TABLE I**

THE PARAMETER ESTIMATES AND THEIR ERRORS

From Table I and Fig. 1, we can draw the conclusions: Increasing data length generally leads to smaller parameter estimation errors; it is clear that the errors \( \delta \) and \( \delta \) are becoming smaller (in general) as \( t \) increases. This confirms the proposed theorem.

**VI. CONCLUSIONS**

An iterative and a recursive algorithms based on replacing unmeasurable noise variables by the estimation residuals are derived for MA models. The analysis using the martingale convergence theorem indicates that the proposed recursive least-squares algorithm of MA models can give consistent parameter estimation. Although the algorithms are developed for the MA models, they can be extended to identify AR models and ARMA models with simple modifications. The MA model least-squares iterative algorithm presented is quite interesting, but its parameter estimation error bounds analysis is more difficult and is worth further research.

**REFERENCES**


