2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

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 ones. 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 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 of work 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 reason: 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. Many correlation analysis based methods are not suitable for identifying the parameters of non-stationary AR, MA and ARMA systems, e.g., [1]–[4], [7], [8]. Therefore, exploring the estimation algorithms and their properties of 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:

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

or

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

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.,

$$(C1)\ \mathrm{E}[v(t)] = 0;\ \mathrm{E}[v(t)v(j)] = 0, t \neq j;\ \mathrm{E}[v^2(t)] = \sigma^2.$$

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 $R(j) := \mathrm{E}[y(t)y(t+j)]$ of the time series $y(t)$ does not depend on t and equals the time-averaged value, i.e.,

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

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

$$\begin{aligned} y(t-j)y(t) &= d_0^2 v(t-j)v(t) + d_0 d_1 v(t-j)v(t-1) \\ &+ d_0 d_2 v(t-j)v(t-2) + d_1 d_0 v(t-j-1)v(t) \\ &+ d_1^2 v(t-j-1)v(t-1) + d_1 d_2 v(t-j-1)v(t-2) \\ &+ d_2 d_0 v(t-j-2)v(t) + d_2 d_1 v(t-j-2)v(t-1) \\ &+ d_2^2 v(t-j-2)v(t-2). \end{aligned}$$

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:

$$\left\{\begin{array}{lcl} d_0^2\sigma^2+d_1^2\sigma^2+d_2^2\sigma^2 & = & R(0), \\ d_0d_1\sigma^2+d_1d_2\sigma^2 & = & R(1), \\ d_0d_2\sigma^2 & = & R(2). \end{array}\right. \quad (4)$$

Then the $n+1=3$ unknown parameters d_1, d_2 and σ^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.,

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

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

$$\left\{\begin{array}{l} d_0^2\sigma^2(t)+d_1^2\sigma^2(t-1)+d_2^2\sigma^2(t-2)=R(t,0), \\ d_0d_1\sigma^2(t-1)+d_1d_2\sigma^2(t-2)=R(t-1,1), \\ d_0d_2\sigma^2(t-2)=R(t-2,2). \end{array}\right. \quad (5)$$

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 θ and information vector $\varphi_0(t)$ as

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

and

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

From (1), we easily get

$$y(t) = \varphi_0^{\mathrm{T}}(t)\theta+v(t), \quad (6)$$

$$Y(t) = \Gamma_0(t)\theta+V(t), \quad (7)$$

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

$$\hat{\theta}=[\Gamma_0^{\mathrm{T}}(t)\Gamma_0(t)]^{-1}\Gamma_0^{\mathrm{T}}(t)Y(t). \quad (8)$$

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 θ at iteration k, then the unknown variable $V(t-i)$ can be computed/estimated by

$$\hat{V}_k(t-i)=Y(t-i)-\hat{\Gamma}_{k-1}(t-i)\hat{\theta}_{k-1},\ i=0,1,\cdots,n \quad (9)$$

with

$$\hat{\Gamma}_k(t):=[\hat{V}_k(t-1),\ \hat{V}_k(t-2),\ \cdots,\ \hat{V}_k(t-n)]\in\mathbb{R}^{p\times n}. \quad (10)$$

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

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

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 $\hat{\Gamma}_0(t)=\mathbf{0}$ 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 ε, if

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

then terminate the procedure and obtain the iterative times k and estimate θ_k; otherwise, increment k by 1 and go to step 3.

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

IV. The recursive algorithm

Let us introduce some notation first. $|X| = \det[X]$ represents the determinant of the matrix X; $\lambda_{\max}[X]$ and $\lambda_{\min}[X]$ 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 δ_1 such that $|f(t)| \leq \delta_1 g(t)$.

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 θ 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 θ based on the noise estimation,

$$\begin{aligned}
&\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)\Gamma^{\mathrm{T}}(t)\cdot \\
&\qquad [Y(t) - \Gamma(t)\hat{\theta}(t-1)], && (12)\\
&P^{-1}(t) = P^{-1}(t-1) + \Gamma^{\mathrm{T}}(t)\Gamma(t), && (13)\\
&\hat{V}(t-i) = Y(t-i) - \Gamma(t-i)\hat{\theta}(t), && (14)\\
&\Gamma(t) = [\hat{V}(t-1),\ \hat{V}(t-2),\ \cdots,\ \hat{V}(t-n)] \\
&\qquad =: [\varphi(t),\ \varphi(t-1),\ \cdots,\ \varphi(t-p+1)]^{\mathrm{T}}. && (15)
\end{aligned}$$

Here, we take $P(0) = p_0 I$ with p_0 being a large positive number, e.g., $p_0 = 10^6$, $\hat{\theta}(0) = \mathbf{1}_n/p_0$, and we refer to $\hat{V}(t)$ as the estimate of $V(t)$.

As $p = 1$, $e(t) := y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1) \in \mathbb{R}^1$ is called the innovation [6], and $E(t) := Y(t) - \Phi^{\mathrm{T}}(t)\hat{\theta}(t-1) \in \mathbb{R}^p$ may be referred to as the innovation vector, i.e., multi-innovation. Thus, 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-MILS algorithm for short.

This MA-MILS 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 θ? The following answers this question.

Define

$$\begin{aligned}
&P_0^{-1}(t) := P_0^{-1}(t-1) + \Gamma_0^{\mathrm{T}}(t)\Gamma_0(t),\ P_0(0) = p_0 I;\\
&r_0(t) := \mathrm{tr}[P_0^{-1}(t)],\ r(t) := \mathrm{tr}[P^{-1}(t)].
\end{aligned}$$

It follows that

$$\begin{aligned}
&|P^{-1}(t)| \leq r^n(t);\ r(t) \geq |P^{-1}(t)|^{1/n};\\
&r(t) \leq n\lambda_{\max}[P^{-1}(t)];\ \ln|P^{-1}(t)| = O(\ln r(t)). \qquad (16)
\end{aligned}$$

In order to establish the main results in this paper, some mathematical preliminaries are required.

Lemma 1: For each i $(i = 0, 1, \cdots, p-1)$, the following inequality holds:

$$\sum_{t=1}^{\infty} \frac{\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)}{[\ln r(t)]^{\beta}} < \infty, \text{ a.s., for any } \beta > 1.$$

Proof From the definition of $P(t)$ in (13), we have

$$\begin{aligned}
P^{-1}(t-1) &= P^{-1}(t) - \Gamma^{\mathrm{T}}(t)\Gamma(t)\\
&\leq P^{-1}(t) - \varphi(t-i)\varphi^{\mathrm{T}}(t-i)\\
&= P^{-1}(t)[I - P(t)\varphi(t-i)\varphi^{\mathrm{T}}(t-i)].
\end{aligned}$$

Taking determinants on both sides and using the formula $\det[I+DE] = \det[I+ED]$ yields

$$\begin{aligned}
|P^{-1}(t-1)| &\leq |P^{-1}(t)||I - P(t)\varphi(t-i)\varphi^{\mathrm{T}}(t-i)|\\
&= |P^{-1}(t)|[1 - \varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)].
\end{aligned}$$

Hence

$$\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i) \leq \frac{|P^{-1}(t)| - |P^{-1}(t-1)|}{|P^{-1}(t)|}.$$

Dividing $[\ln r(t)]^{\beta}$ and summing for t give (noting that $|P^{-1}(t)|$ is a non-decreasing function of t)

$$\begin{aligned}
&\sum_{t=1}^{\infty} \frac{\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)}{[\ln r(t)]^{\beta}} \leq n^{\beta} \sum_{t=1}^{\infty} \frac{\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)}{[\ln |P^{-1}(t)|]^{\beta}}\\
&\leq n^{\beta} \sum_{t=1}^{\infty} \frac{|P^{-1}(t)| - |P^{-1}(t-1)|}{|P^{-1}(t)|[\ln |P^{-1}(t)|]^{\beta}}\\
&= n^{\beta} \sum_{t=1}^{\infty} \int_{|P^{-1}(t-1)|}^{|P^{-1}(t)|} \frac{\mathrm{d}x}{|P^{-1}(t)|[\ln |P^{-1}(t)|]^{\beta}}\\
&\leq n^{\beta} \sum_{t=1}^{\infty} \int_{|P^{-1}(t-1)|}^{|P^{-1}(t)|} \frac{\mathrm{d}x}{x(\ln x)^{\beta}}\\
&\leq n^{\beta} \int_{|P^{-1}(0)|}^{|P^{-1}(\infty)|} \frac{\mathrm{d}x}{x(\ln x)^{\beta}} = \frac{-n^{\beta}}{\beta-1} \left.\frac{1}{(\ln x)^{\beta-1}}\right|_{|P^{-1}(0)|}^{|P^{-1}(\infty)|}\\
&= \frac{n^{\beta}}{\beta-1}\left(\frac{1}{[\ln |P^{-1}(0)|]^{\beta-1}} - \frac{1}{[\ln |P^{-1}(\infty)|]^{\beta-1}}\right) < \infty, \text{ a.s.}
\end{aligned}$$

$\square$

Define the parameter estimation error vector $\tilde{\theta}(t)$ and a nonnegative definite function $W(t)$ as

$$\begin{aligned}
\tilde{\theta}(t) &= \hat{\theta}(t) - \theta, && (17)\\
W(t) &= \tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t)\tilde{\theta}(t). && (18)
\end{aligned}$$

Lemma 2: For the system in (6) and the MA-MILS algorithm in (12)-(15), assume the noise sequence $\{v(t)\}$ with zero mean and bounded time-varying variance satisfies [3]:

$$\begin{aligned}
&(A1) \qquad \mathrm{E}[v(t)|\mathscr{F}_{t-1}] = 0, \text{ a.s.,}\\
&(A2) \qquad \mathrm{E}[v^2(t)|\mathscr{F}_{t-1}] = \sigma_v^2(t) \leq \bar{\sigma}_v^2 < \infty, \text{ a.s.,}\\
&(A3) \qquad H(z) = \frac{1}{D(z)} - \frac{1}{2} \text{ is strictly positive real,}
\end{aligned}$$

where $\{v(t), \mathscr{F}_t\}$ is a martingale sequence defined on a probability space $\{\Omega, \mathscr{F}, P\}$ and $\{\mathscr{F}_t\}$ is the σ algebra sequence generated by $\{v(t)\}$. Then the following inequality holds:

$$\begin{aligned}
\mathrm{E}[W(t) + S(t)|\mathscr{F}_{t-1}] &\leq W(t-1) + S(t-1)\\
&\quad + 2p\sum_{i=0}^{p-1} \varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)\bar{\sigma}_v^2, \text{ a.s.,}
\end{aligned}$$

where

$$\begin{aligned} S(t) &:= 2\sum_{i=1}^{t}\tilde{U}^{\mathrm{T}}(i)\tilde{Y}(i), \text{ a.s.}, \\ \tilde{Y}(t) &:= \frac{1}{2}\Gamma(t)\tilde{\theta}(t)+[Y(t)-\Gamma(t)\hat{\theta}(t)-V(t)], \quad (19)\\ \tilde{U}(t) &:= -\Gamma(t)\tilde{\theta}(t). \quad (20)\end{aligned}$$

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 $\hat{V}(t)$ as follows:

$$E(t) := Y(t)-\Gamma(t)\hat{\theta}(t-1), \quad (21)$$

$$\hat{V}(t) := Y(t)-\Gamma(t)\hat{\theta}(t). \quad (22)$$

It follows that

$$\begin{aligned}\hat{V}(t) &= [I-\Gamma(t)P(t)\Gamma^{\mathrm{T}}(t)]E(t)\\ &= [I+\Gamma(t)P(t-1)\Gamma^{\mathrm{T}}(t)]^{-1}E(t). \quad (23)\end{aligned}$$

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

$$\begin{aligned}\tilde{\theta}(t) &= \tilde{\theta}(t-1)+P(t)\Gamma^{\mathrm{T}}(t)E(t)\\ &= \tilde{\theta}(t-1)+P(t-1)\Gamma^{\mathrm{T}}(t)\hat{V}(t), \quad (24)\end{aligned}$$

or

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

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

$$\begin{aligned}\tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t-1)\tilde{\theta}(t) &= [\tilde{\theta}(t-1)+P(t-1)\Gamma^{\mathrm{T}}(t)\hat{V}(t)]^{\mathrm{T}}\cdot\\ &\quad P^{-1}(t-1)\tilde{\theta}(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t).\end{aligned}$$

Using (13), we have

$$\begin{aligned}\tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t)\tilde{\theta}(t) &= \tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t)\\ &\quad +\tilde{\theta}^{\mathrm{T}}(t-1)P^{-1}(t-1)\tilde{\theta}(t-1)\\ &\quad +\hat{V}^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t).\end{aligned}$$

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

$$\begin{aligned} W(t) &= W(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t)\\ &\quad +\hat{V}^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t).\\ &= W(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t)\\ &\quad +\hat{V}^{\mathrm{T}}(t)\Gamma(t)[\tilde{\theta}(t)-P(t)\Gamma^{\mathrm{T}}(t)E(t)]\\ &\quad +\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t)\\ &= W(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t)\\ &\quad +2\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t)-\hat{V}^{\mathrm{T}}(t)\Gamma(t)P(t)\Gamma^{\mathrm{T}}(t)E(t)\\ &= W(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t)\\ &\quad +2\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t)\\ &\quad -E^{\mathrm{T}}(t)[I-\Gamma(t)P(t)\Gamma^{\mathrm{T}}(t)]\Gamma(t)P(t)\Gamma^{\mathrm{T}}(t)E(t)\\ &\leq W(t-1)+\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\Gamma(t)\tilde{\theta}(t)\\ &\quad +2\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)\hat{V}(t)\\ &= W(t-1)+2\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)V(t)\\ &\quad +2\tilde{\theta}^{\mathrm{T}}(t)\Gamma^{\mathrm{T}}(t)[\frac{1}{2}\Gamma(t)\tilde{\theta}(t)+\hat{V}(t)-V(t).\end{aligned}$$

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

$$\begin{aligned} W(t) &\leq W(t-1)-2\tilde{U}^{\mathrm{T}}(t)\tilde{Y}(t)\\ &\quad +2[\tilde{\theta}(t-1)+P(t)\Gamma^{\mathrm{T}}(t)E(t)]^{\mathrm{T}}\Gamma^{\mathrm{T}}(t)V(t)\\ &= W(t-1)-2\tilde{U}^{\mathrm{T}}(t)\tilde{Y}(t)\\ &\quad +2\tilde{\theta}^{\mathrm{T}}(t-1)\Gamma^{\mathrm{T}}(t)V(t)\\ &\quad +2[E(t)-V(t)]^{\mathrm{T}}\Gamma(t)P(t)\Gamma^{\mathrm{T}}(t)V(t)\\ &\quad +2V^{\mathrm{T}}(t)\Gamma(t)P(t)\Gamma^{\mathrm{T}}(t)V(t). \quad (26)\end{aligned}$$

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

$$\begin{aligned}\mathrm{E}[W(t)|\mathscr{F}_{t-1}] &\leq W(t-1)-2\mathrm{E}[\tilde{U}^{\mathrm{T}}(t)\tilde{Y}(t)|\mathscr{F}_{t-1}]\\ &\quad +2p\sum_{i=0}^{p-1}\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)\bar{\sigma}_v^2, \text{ a.s.} \quad (27)\end{aligned}$$

Since

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

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

$$\begin{aligned}\tilde{Y}(t) &= \frac{1}{2}\Gamma(t)\tilde{\theta}(t)+[\hat{V}(t)-V(t)]\\ &= \left[\frac{1}{D(z)}-\frac{1+\rho}{2}\right]\tilde{U}(t)+\frac{\rho}{2}\tilde{U}(t)\\ &=: \tilde{Y}_1(t)+\frac{\rho}{2}\tilde{U}(t),\end{aligned}$$

where

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

Here $\tilde{Y}_1(t)$ 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

$$2\sum_{i=1}^{t}\tilde{U}^{\mathrm{T}}(i)\tilde{Y}_1(i)\geq 0, \text{ a.s.},$$

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

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:

$$\|\hat{\theta}(t)-\theta\|^2=O\left(\frac{[\ln r_0(t)]^{\beta}}{\lambda_{\min}[P_0^{-1}(t)]}\right), \text{ a.s.}$$

Proof From the definition of $W(t)$, we have

$$\|\tilde{\theta}(t)\|^2 \leq \frac{\tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t)\tilde{\theta}(t)}{\lambda_{\min}[P^{-1}(t)]} = \frac{W(t)}{\lambda_{\min}[P^{-1}(t)]}. \tag{30}$$

Let

$$Z(t) = \frac{W(t)+S(t)}{[\ln r(t)]^{\beta}}.$$

Since $\ln r(t)$ is non-decreasing, according to Lemma 2, we have

$$\begin{aligned}\mathrm{E}[Z(t)|\mathscr{F}_{t-1}] &\leq \frac{W(t-1)+S(t-1)}{[\ln r(t)]^{\beta}} \\ &\quad +2p\sum_{i=0}^{p-1}\frac{\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)}{[\ln r(t)]^{\beta}}\,\bar{\sigma}_v^2 \\ &\leq Z(t-1)+2p\sum_{i=0}^{p-1}\frac{\varphi^{\mathrm{T}}(t-i)P(t)\varphi(t-i)}{[\ln r(t)]^{\beta}}\,\bar{\sigma}_v^2. \end{aligned} \tag{31}$$

Using Lemma 1, it is clear that the sum for t from 1 to ∞ 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.,

$$Z(t) = \frac{W(t)+S(t)}{[\ln r(t)]^{\beta}} \to Z_0 < \infty, \text{ a.s.},$$

or

$$W(t) = O([\ln r(t)]^{\beta}), \text{ a.s., } S(t) = O([\ln r(t)]^{\beta}), \text{ a.s.} \tag{32}$$

Since $H(z)$ is a strictly positive real function, from (29), it follows that

$$\sum_{i=1}^{t}\|\tilde{U}(i)\|^2 = O([\ln r(t)]^{\beta}).$$

From (30), (32), we have

$$\|\tilde{\theta}(t)\|^2 = O\left(\frac{[\ln r(t)]^{\beta}}{\lambda_{\min}[P^{-1}(t)]}\right), \text{ a.s., for any } \beta > 1.$$

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

$$\begin{aligned}\sum_{i=1}^{t}\|\hat{V}(i)-V(i)\|^2 &\leq k_1\sum_{i=1}^{t}\|\tilde{U}(i)\|^2+k_2 \\ &= O([\ln r(t)]^{\beta}).\end{aligned}$$

Now we prove $r(t) = O(r_0(t))$, $\lambda_{\min}[P^{-1}(t)] = O(\lambda_{\min}[P_0^{-1}(t)]$. Define the vector error $\tilde{\Gamma}(t)$ as follows [refer to the definitions of $\Gamma(t)$ and $\Gamma_0(t)$]:

$$\begin{aligned}\tilde{\Gamma}(t) &:= \Gamma(t)-\Gamma_0(t) \\ &= [\hat{V}(t-1)-V(t-1),\cdots,\hat{V}(t-n)-V(t-n)].\end{aligned}$$

Hence, for any $\beta > 1$, we have

$$\begin{aligned}\sum_{i=1}^{t}\|\tilde{\Gamma}(i)\|^2 &= \sum_{i=1}^{t}\sum_{j=1}^{n}\|\hat{V}(i-j)-V(i-j)\|^2 \\ &= O([\ln r(t)]^{\beta}), \\ r(t) &\leq 2r_0(t)+2\sum_{i=1}^{t}\|\tilde{\Gamma}(i)\|^2 \\ &= 2r_0(t)+O([\ln r(t)]^{\beta}) = O(r_0(t)), \text{ a.s.}\end{aligned}$$

For any vector $\omega \in \mathbb{R}^n$ with $\|\omega\| = 1$, we have

$$\begin{aligned}\sum_{i=1}^{t}\|\Gamma(i)\omega\|^2 &= \sum_{i=1}^{t}\|\Gamma_0(i)\omega-\tilde{\Gamma}(i)\omega\|^2 \\ &\leq 2\sum_{i=1}^{t}\|\Gamma_0(i)\omega\|^2+2\sum_{i=1}^{t}\|\tilde{\Gamma}(i)\|^2 \\ &= 2\sum_{i=1}^{t}\|\Gamma_0(i)\omega\|^2+O([\ln r_0(t)]^{\beta}).\end{aligned}$$

Thus

$$\begin{aligned}\lambda_{\min}[P^{-1}(t)] &\leq 2\lambda_{\min}[P_0^{-1}(t)]+O(\lambda_{\min}[P_0^{-1}(t)]) \\ &= O(\lambda_{\min}[P_0^{-1}(t)]).\end{aligned}$$

Hence, it is not difficult to get

$$\|\hat{\theta}(t)-\theta\|^2 = O\left(\frac{[\ln r_0(t)]^{\beta}}{\lambda_{\min}[P_0^{-1}(t)]}\right), \text{ a.s., for any } \beta > 1.$$

This proves Theorem 1. □

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:

$$(C1) \qquad c_1 I \leq \frac{1}{t}\sum_{i=1}^{t}\Gamma_0(i)\Gamma_0^{\mathrm{T}}(i) \leq c_2 t^{c_0} I, \text{ a.s.}$$

Then for any $\beta > 1$, we have

$$\|\hat{\theta}(t)-\theta\|^2 = O\left(\frac{[\ln t]^{\beta}}{t}\right) \to 0, \text{ a.s.}$$

Since $\ln t = o(t^{\varepsilon})$ ($\varepsilon > 0$: arbitrary small), $\|\hat{\theta}(t)-\theta\|$ converges to zero approximately at the rate of $1/\sqrt{t^{1-\varepsilon}}$ 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,

$$\begin{aligned}\hat{\theta}(t) &= \hat{\theta}(t-1)+P(t)\varphi(t)[y(t)-\varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)], \\ P(t) &= P(t-1)-\frac{P(t-1)\varphi(t)\varphi^{\mathrm{T}}(t)P(t-1)}{1+\varphi^{\mathrm{T}}(t)P(t-1)\varphi(t)}, \\ \hat{v}(t) &= y(t)-\varphi^{\mathrm{T}}(t)\hat{\theta}(t), \\ \varphi(t) &= [\hat{v}(t-1),\ \hat{v}(t-2),\ \cdots,\ \hat{v}(t-n)]^{\mathrm{T}} \in \mathbb{R}^n.\end{aligned}$$

V. Example

An example is given to demonstrate the effectiveness of the proposed algorithm. Consider the following simulation plant:

$$\begin{aligned}y(t) &= D(z)v(t), \\ D(z) &= 1+d_1z^{-1}+d_2z^{-2} \\ &= 1+0.412z^{-1}+0.309z^{-2}, \\ \theta &= [d_1,\ d_2]^{\mathrm{T}} = [0.412,\ 0.309]^{\mathrm{T}}.\end{aligned}$$

$\{v(t)\}$ 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 and their errors are shown in Table I, and the parameter estimation error δ versus t is shown in Fig. 1, where $\delta = \|\hat{\theta}(t) - \theta\|/\|\theta\|$ is the relative parameter estimation errors.

TABLE I

THE PARAMETER ESTIMATES AND THEIR ERRORS

t	d_1	d_2	δ (%)
100	0.54958	0.53833	51.92873
200	0.52710	0.43024	32.46150
300	0.44045	0.36930	12.94705
500	0.41298	0.35975	9.85596
800	0.42472	0.36475	11.10289
1000	0.43643	0.37015	12.78600
1500	0.42497	0.35508	9.29554
2000	0.41717	0.36141	10.22612
2500	0.41748	0.34951	7.93853
3000	0.40954	0.34563	7.12897
3500	0.40916	0.33302	4.69745
4000	0.41801	0.32058	2.53334
4500	0.41634	0.31570	1.54975
5000	0.41741	0.31518	1.59455
5500	0.42543	0.30820	2.61283
6000	0.41706	0.31131	1.08043
6500	0.41579	0.31373	1.17648
7000	0.41490	0.31742	1.72851
7500	0.41181	0.31066	0.32461
8000	0.40790	0.31275	1.07837
8500	0.41006	0.31346	0.94452
9000	0.41200	0.31338	0.84971
9500	0.41068	0.31055	0.39596
10000	0.40994	0.30678	0.58897
True values	0.41200	0.30900	

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 δ and δ 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

[1] F. Desbouvries, I. Fijalkow, and P. Loubaton "On the identification of noisy MA models," *IEEE Trans. Automat. Contr.*, vol. 41, no. 12, pp. 1810-1814, 1996.

[2] J. Franke "A Levinson-Durbin recursion for autoregressive-moving average processes," *Biometrika*, vol. 72, no. 3, pp. 573-581, 1985.

[3] G.C. Goodwin and K.S. Sin, *Adaptive Filtering, Prediction and Control*, Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

[4] D. Graupe, *Time series analysis, Identification and Adaptive Filtering*, Robert E. Krieger Publishing Company, Inc., Florida, USA, 1984.

[5] D. Graupe, D.J. Krause, and J.B. Moore "Identification of autoregressive moving-average parameters of time series," *IEEE Trans. Automat. Contr.*, vol. 20, no. 1, pp. 104-107, 1975.

[6] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1999.

[7] W. Wu and P. Chen, "Adaptive AR modeling in white Gaussian noise," *IEEE Trans. Signal Processing*, vol. 45, no. 5, pp. 1184-1192, 1997.

[8] S. Li, Y. Zhu, and B.W. Dickinson, "A comparison of two linear methods of estimating the parameters of ARMA models," *IEEE Trans. Automat. Contr.*, vol. 34, no. 8, pp. 915-917, 1989.

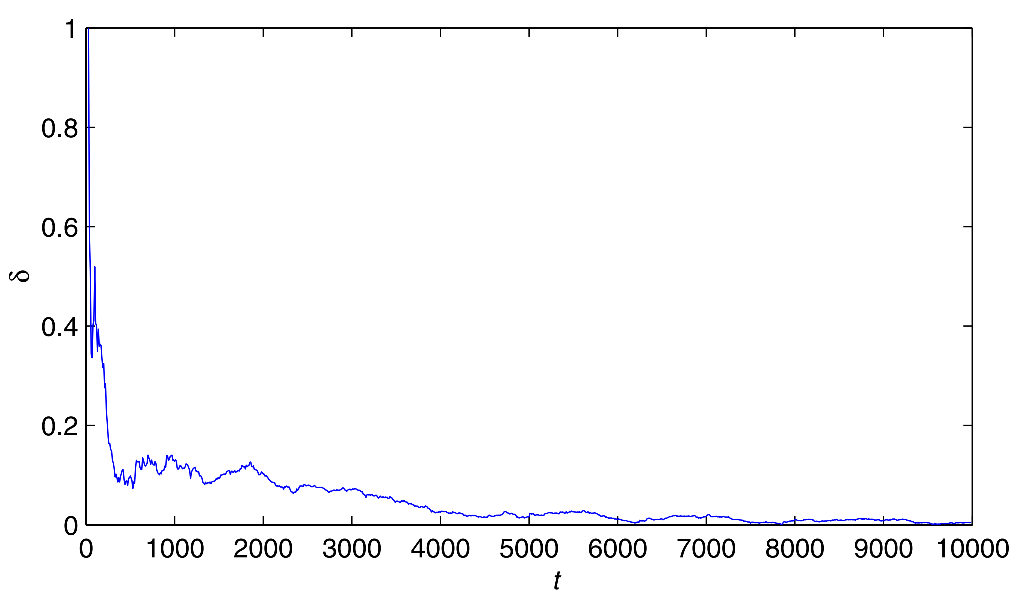

Fig. 1. The parameter estimation errors δ vs. t