PARAMETER IDENTIFICATION
FOR
STOCHASTIC SYSTEMS

Junko MURAKAMI

Adviser
Dr. Bozenna PASIK-DUNCAN

Department of Mathematics,
University of Kansas,
Lawrence, U.S.A.

Spring 1995

1. Introduction

This paper discusses several methods for identifying the values of unknown parameters in a linear stochastic system; i.e., the maximum likelihood method, the least squares method, the extended least squares method, the weighted least squares method, and the prediction error method. The estimators obtained will be checked for convergence, and their variances will be discussed. Also, whether they are biased will also be checked. For each method, some examples and the computer simulation of the process will be shown.

Consider the linear stochastic system

$$x_{k+1} = A\,x_k + B\,u_k + G\,w_k, \tag{1}$$

$$y_k = C\,x_k + H\,v_k$$

where $x_k \in R^p$, $u_k \in R^m$, $y_k \in R^l$, $w_k \in R^g$, $v_k \in R^h$ and A, B, G, C, H are fixed matrices of appropriate dimension. The basic random variables x_0, w_0, w_1, $\ldots$, v_0, v_1, $\ldots$ are assumed to be independent. Also, consider the single input single output pth order ARMAX model,

$$y_k + \sum_{i=1}^{p} a_i\,y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + \varepsilon_k + \sum_{i=1}^{p} c_i\,\varepsilon_{k-i}, \tag{2}$$

where $\{y_k\}$ and $\{u_k\}$ are output and input sequences, respectively. The system described in Equation (1) above can be written using this ARMAX model as follows [1].

We introduce the shift operator q. For any sequence or stochastic process $\{\xi_k,\ k = 0, 1, \ldots\}$, let ξ be the sequence $\{\xi_k\}$, and let $q\xi$ be the sequence $\{q\xi_k\}$ where $q\xi_k := \xi_{k+1}$. Then (1) can be written as

$$qx = A\,x + B\,u + G\,w, \qquad \text{(i)}$$

$$y = C\,x + H\,v, \qquad \text{(ii)}$$

where x, y, u, v, and w denote sequences. Solve (i) for x,

$$(qI - A)x = B\,u + G\,w$$

$$x = (qI - A)^{-1}[B\,u + G\,w],$$

and substitute x in (ii) with the equation above, we get

$$y = C(qI - A)^{-1}[B\,u + G\,w] + H\,v. \qquad \text{(iii)}$$

Since adj$(qI - A)$ is a polynomial of q with degree $p-1$ and with the coefficients in $p\times p$ matrices, we can let α_i be matrices such that adj$(qI - A) = \alpha_p + \alpha_{p-1}q + \cdots + \alpha_1 q^{p-1}$. Note

$$(qI - A)^{-1} = [\det(qI - A)]^{-1}\mathrm{adj}(qI - A)$$

and

$$\det(qI - A) = (qI - A)[\mathrm{adj}(qI - A)].$$

Also,

$$\begin{aligned}\det(qI - A) &= (qI - A)[\mathrm{adj}(qI - A)] \\ &= (qI - A)(\alpha_p + \alpha_{p-1}q + \cdots + \alpha_1 q^{p-1}) \\ &= \alpha_p q + \alpha_{p-1}q^2 + \cdots + \alpha_2 q^{p-1} + \alpha_1 q^p - A\alpha_p - A\alpha_{p-1}q - \cdots - A\alpha_1 q^{p-1} \\ &= -A\alpha_p + (\alpha_p - A\alpha_{p-1})q + (\alpha_{p-1} - A\alpha_{p-2})q^2 + \\ &\qquad \cdots + (\alpha_2 - A\alpha_1)q^{p-1} + \alpha_1 q^p \\ &= -A\alpha_p + \sum_{i=1}^{p-1}(\alpha_{p+1-i} - A\alpha_{p-i})q^i + \alpha_1 q^p.\end{aligned}$$

Then

$$\begin{aligned}q^{-p}[\det(qI - A)] &= -A\alpha_p\, q^{-p} + \sum_{i=1}^{p-1}(\alpha_{p+1-i} - A\alpha_{p-i})q^{i-p} + \alpha_1 \\ &= \sum_{i=0}^{p} a_i q^{-i}\end{aligned}$$

where $a_0 = I$, $a_p = -A\alpha_p$, and $a_i = \alpha_{i+1} - A\alpha_i$ for $1 \le i \le p-1$. Multiply (iii) above by $q^{-p}[\det(qI - A)]$, and we get

$$q^{-p}[\det(qI - A)]y = q^{-p}[\det(qI - A)]C(qI - A)^{-1}[B\,u + G\,w] + q^{-p}[\det(qI - A)]H\,v.$$

$$
\begin{aligned}
[\sum_{i=0}^{p} a_i q^{-i}]y &= q^{-p}[\det(qI - A)]C \{[\det(qI - A)]^{-1}\mathrm{adj}(qI - A)\} [B\, u + G\, w] \\
&\qquad + [\sum_{i=0}^{p} a_i q^{-i}]H\, v \\
&= q^{-p}C\, \mathrm{adj}(qI - A)[B\, u + G\, w] + [\sum_{i=0}^{p} a_i q^{-i}]H\, v \\
&= q^{-p}C[\sum_{i=1}^{p} \alpha_i q^{p-i}][B\, u + G\, w] + [\sum_{i=0}^{p} a_i q^{-i}]H\, v \\
&= [\sum_{i=1}^{p} C\alpha_i q^{-i}][B\, u + G\, w] + [\sum_{i=0}^{p} a_i q^{-i}]H\, v \\
&= [\sum_{i=1}^{p} C\alpha_i B q^{-i}]u + [\sum_{i=1}^{p} C\alpha_i G q^{-i}]w + [\sum_{i=0}^{p} a_i H q^{-i}]v \\
&= [\sum_{i=1}^{p} C\alpha_i B q^{-i}]u + [0, H]\begin{bmatrix} w \\ v \end{bmatrix} + \sum_{i=1}^{p} [C\alpha_i \mathrm{G},\, a_i H]\, q^{-i}\begin{bmatrix} w \\ v \end{bmatrix} \\
&= [\sum_{i=1}^{p} b_i q^{-i}]u + c_0 \varepsilon + [\sum_{i=1}^{p} c_i q^{-i}]\varepsilon
\end{aligned}
$$

where $\varepsilon = [w, v]^T$, $b_i = C\alpha_i B$, $c_0 = [0, H]$, and $c_i = [C\alpha_i G, a_i H]$ for $1 \leq i \leq p$ [1]. So, we can write

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + c_0 \varepsilon_k + \sum_{i=1}^{p} c_i \varepsilon_{k-i}.$$

This is the ARMAX model, where $y_k \in R^l$, $u_k \in R^m$, $\varepsilon_k \in R^{g+h}$. Also, under the condition that $H = I$, this model can be written as Equation (2).

By choosing an appropriate substitution for θ^o, ϕ_k^T, and v—for example, letting $\theta^o := (-a_1, \ldots, -a_p, b_1, \ldots, b_p)^T$, $\phi_k^T := (y_k, \ldots, y_{k-p+1}, u_k, \ldots, u_{k-p})$, and $w_{k+1} := (c_1\varepsilon_k + \cdots + c_p\varepsilon_{k-p+1} + \varepsilon_{k+1})$, etc.—we can rewrite Equation (2) above as follows:

$$y_{k+1} = \phi_k^T \theta^o + w_{k+1} \tag{3}$$

where y_{k+1} and ϕ_k are random variables linearly related through the parameter θ^o, except for the perturbation w_{k+1}. Our goal is to identify the parameter θ^o.

Some of the terminologies that are critical to an identification are described below.

Markov Chain

Let $\{X_n, n = 0, 1, 2, \ldots\}$ be a stochastic process that takes on a finite or countable number of possible values, where n is the set of nonnegative integers. If $X_n = i$, then the process is said to be in state i at time n. Suppose there exists P_{ij} denoting a fixed probability that the process currently in state i will next be in state j; i.e., we suppose that

$$P_{ij} = P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\}$$

exists for all states $i_0, i_1, \ldots, i_{n-1}, i, j$ and all $n \geq 0$. Such a stochastic process is known as a **Markov chain**. Therefore, since the above probability depends only on the current state i and the next state j, we can write:

$$P_{ij} = P\{X_{n+1} = j | X_n = i\}.$$

In other words, the conditional distribution of any future state X_{n+1}, given the past states X_0, X_1, ..., X_{n-1} and the present state X_n, is independent of the past states and depends only on the present state. This property is called a **Markovian property**.

Martingale Process

A stochastic process $\{X_k, F_k, k \geq 1\}$ is said to be a **martingale** process

i) if F_k is an increasing sequence of σ-algebras,
ii) if X_k is F_k-measurable, and
iii) if $E\,(X_{k+1} \mid F_k) = X_k$ a.s. for all k.

The Martingale Convergence Theorem:

If $\{X_k, F_k, k \geq 1\}$ is a martingale such that for some $p \geq 1$,

$$\sup_k E\,(|X_k|^p) < \infty,$$

then $\lim_{k\to\infty} X_k$ exits and is finite a.s.

Kronecker's Lemma

Let $\{r_k\}$ and $\{s_k\}$ be two real valued sequences satisfying

$$r_k > 0, \;\; \lim_{k\to\infty} r_k = \infty, \;\text{ and } \sum_{k=1}^{\infty} \frac{s_k}{r_k} < \infty.$$

Then

$$\lim_{N\to\infty} \frac{1}{r_N} \sum_{k=1}^{N} s_k = 0.$$

Brownian Motion Process (Wiener Process)

Let $\{X(t),\ t \in T\}$ be a stochastic process, where T is the index set. The process has *independent increments* if for all $t_0 < t_1 < t_2 < \cdots < t_n$, the random variables

$$X(t_1) - X(t_0),\ X(t_2) - X(t_1),\ \ldots,\ X(t_n) - X(t_{n-1})$$

are independent. Furthermore, it has *stationary increments* if $X(t + s) - X(t)$ has the same distribution for all t.

A stochastic process $[W(t),\ t \geq 0]$ is said to be a **Brownian motion process** if:

(i) $W(0) = 0$;
(ii) $\{W(t),\ t \geq 0\}$ has stationary independent increments;
(iii) for every $t > 0$, $W(t)$ is normally distributed with mean 0 and variance c^2t.

If $c = 1$, the process is called **standard Brownian motion.**

2. Estimators

Consider a system given in Equation (2) which has input u and output y. There is a conditional probability distribution

$$F(z \mid u) := P\{y \leq z \mid u\}$$

of y given u. Since F depends on the parameter θ, we perform some experiment on the system, that is we apply an input u and observe the resulting output y, and make an estimate $\hat{\theta} = \tau(y,u) \in R^q$ for the true parameter θ^o. The function τ is called the **estimator** and $\hat{\theta}$ is the **estimate.**

There are two approaches to find an estimator: the Bayesian approach and the non-Bayesian approach. Both of the approaches will be discussed. However, we will focus our attention mainly to the non-Bayesian approach, which includes the **maximum likelihood estimator (MLE)**, the **least squares estimator (LSE)**, the **extended least squares estimator (ELS)**, the **weighted least squares estimator (WLS)**, and the **prediction error estimator (PEE)**.

Clearly, we need to make sure that we have the estimator that converges. Furthermore, having the estimator in recursive form is very important for the actual computation, as discussed below.

2.1 The Bayesian Approach

In the Bayesian approach we assume to know $p(\theta)$, on the parameter set Θ, which represents our belief of the likelihood that $\theta = \theta^o$ before making any observations. From this prior density function $p(\theta)$ we can obtain the conditional density function $p(\theta \mid y, u)$ by Bayes' rule,

$$\begin{aligned} p(\theta \mid y, u) &= \frac{p(y \mid u, \theta)\, p(\theta)}{p(y \mid u)} \\ &= \frac{p(y \mid u, \theta)\, p(\theta)}{\int p(y \mid u, \theta)\, p(\theta)\, d\theta} . \end{aligned}$$

2.1.1 The Bayes Estimator

Let $L(\theta', \theta)$ be a *cost function* representing the loss incurred by estimating the parameter to be θ' when $\theta = \theta^o$. Let τ be any estimator. Then the expected loss incurred by τ given (y, u) is

$$J(\tau \mid y, u) := \int_{\Theta} L(\tau(y, u), \theta)\, p(\theta \mid y, u)\, d\theta$$

The **Bayes estimator** τ^* is the one which minimizes the expected loss, i.e.,

$$J(\tau^* \mid y, u) \leq J(\tau \mid y, u), \quad \text{for all } \tau$$

[1].

Example: A communication channel in which the unknown transmitted signal θ is either 0 or 1, so $\theta \in \Theta = (0, 1)$. The received signal is

$$y = \theta + w,$$

where w is the channel noise and $w \sim N(0, \sigma^2)$. The receiver observes y and has to decide which of the two possible signals θ has transmitted. To the receiver's decision is represented by a function τ: $y \rightarrow (0,1)$. Suppose the loss function is $L(\theta, \theta^o) = |\theta - \theta^o|^2$ where θ^o is the true signal transmitted. The prior probability of the transmitted signal is $p_0 = P\{\theta = 0\}$, $p_1 = P\{\theta = 1\} = 1 - p_0$.

By Bayes' rule,

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y \mid 1)p_1 + p(y \mid 0)p_0}$$

and, since $y \sim N(\theta, \sigma)$,

$$p(y \mid \theta) = (\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{(x - \theta)^2}{2\sigma^2}\right].$$

So,

$$p(1 \mid y) = \frac{(\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]p_1}{(\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]p_1 + (\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{x^2}{2\sigma^2}\right]p_0}$$

$$= \frac{\exp\left[-\frac{x^2}{2\sigma^2} - \frac{2x+1}{2\sigma^2}\right]p_1}{\exp\left[-\frac{x^2}{2\sigma^2} - \frac{2x+1}{2\sigma^2}\right]p_1 + \exp\left[-\frac{x^2}{2\sigma^2}\right]p_0}$$

$$= \frac{p_1}{p_1 + p_0\exp[(2x+1)/(2\sigma^2)]}$$

Similarly,

$$p(0 \mid y) = \frac{(\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{x^2}{2\sigma^2}\right]p_1}{(\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]p_1 + (\sigma^2 2\pi)^{-1/2}\exp\left[-\frac{x^2}{2\sigma^2}\right]p_0}$$

$$= \frac{p_0}{p_1\exp[(-2x-1)/(2\sigma^2)] + p_0}.$$

So, let $\alpha = \int_{-t}^{t} p(1 \mid y)dx$, then

$$\alpha = \int_{-t}^{t} \frac{p_1}{p_1 + p_0 \exp[(2x + 1)/(2\sigma^2)]} dx$$
$$= \left[\frac{2x + 1}{2} - \frac{1}{\sigma^2} ln\left(p_1 + p_0 \exp\left[\frac{2x + 1}{2\sigma^2} \right] \right) \right]_{-t}^{t}$$
$$= 2t - \frac{1}{\sigma^2} ln \frac{p_1 + p_0 \exp[(2t + 1)/(2\sigma^2)]}{p_1 + p_0 \exp[(-2t + 1)/(2\sigma^2)]}$$

Similarly, let $\beta = \int_{-t}^{t} p(0 \mid y)dx$, then

$$\beta = \int_{-t}^{t} \frac{p_0}{p_0 + p_1 \exp[-2x - 1)/(2\sigma^2)]} dx$$
$$= -2t + \frac{1}{\sigma^2} ln \frac{p_0 + p_1 \exp[(-2t - 1)/(2\sigma^2)]}{p_0 + p_1 \exp[(2t - 1)/(2\sigma^2)]} .$$

The expected loss is

$$J(\theta \mid y \leq |t|) = (\theta - 1)^2 \int_{-t}^{t} p(1 \mid y)dx + \theta^2 \int_{-t}^{t} p(0 \mid y)dx.$$
$$= (\theta - 1)^2 \alpha + \theta^2 \beta$$

So, the estimator θ that minimizes the above expected loss is

$$\frac{J}{\partial \theta}(\theta \mid y \leq |t|) = 2(\theta - 1)\alpha + 2\theta\beta = 0$$
$$(\alpha + \beta)\theta = \alpha$$
$$\hat{\theta} = \frac{\alpha}{\alpha + \beta} .$$

2.2. The Non-Bayesian Approach

If there is no prior distribution over the set of model parameters Θ, we look for estimators with some desirable properties as shown below.

2.2.1 The Maximum Likelihood Estimator (MLE)

Consider random variables $X_1, X_2, \ldots, X_n$ from a distribution having p.d.f. $g(x;\theta)$, $\theta \in \Theta$. Then, their joint probability function

$$M(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta), \quad \theta \in \Theta$$

which represents the likelihood that the values $x_1, x_2, \ldots, x_n$ will be observed when θ is the true value of the parameter, is called the likelihood function M. Suppose there exists $\hat{\theta}$ such that, when θ is replaced by $\hat{\theta}$, M is a maximum, then $\hat{\theta}$ is called a **maximum likelihood estimator** of θ.

If we consider the model given in Equation (3)

$$y_{k+1} = \phi_k^T \theta + w_{k+1},$$

we see $x_s = (\phi_s, y_s)$ and the likelihood function M is equal to

$$p^\theta(\phi_s, y_s, s \leq n) = p^\theta_{0|-1}(\phi_0) q^\theta_{1|0}(y_1 | \phi_0) p^\theta_{1|0}(\phi_1 \mid \phi_0, y_1) q^\theta_{2|1}(y_2 \mid \phi_0, \phi_1, y_1) \cdots p^\theta_{n|n-1}(\phi_n \mid \phi_0, \ldots, \phi_{n-1}, y_1, \ldots, y_n)$$

where $q^\theta_{k|k-1}$ and $p^\theta_{k|k-1}$ are conditional probability densities.

Consider the model given by Equation (3), which is

$$y_{k+1} = \phi_k^T \theta + w_k + 1.$$

2.2.1.a Suppose the following assumptions hold:

Assumption 1: $\{w_k\}$ is independently identically distributed (iid) with known probability density $f(w)$.

Assumption 2: ϕ_k is independent of $\{w_s, s \geq k+1\}$ for $k = 0, 1, \ldots$

Assumption 3: $p^{\,\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ.

Then the MLE $\hat{\theta}_n$ maximizes

$$M_n(\theta) = q^\theta_{1|0}(y_1 \mid \phi_0) q^\theta_{2|1}(y_2 \mid \phi_0, \phi_1, y_1) \cdots q^\theta_{n|n-1}(y_n \mid \phi_0, \ldots, \phi_{n-1}, y_1, \ldots, y_{n-1}).$$

$$= f(y_1 - \phi_0^T \theta) f(y_2 - \phi_1^T \theta) \cdots f(y_n - \phi_{n-1}^T \theta).$$

So, the MLE maximizes

$$\ln M_n(\theta) = \sum_{k=0}^{n-1} \ln f(y_{k+1} - \phi_k^T \theta).$$

Furthermore, if we change Assumption 1 to Assumption 4 as below, the MLE coincides with the LSE.

2.2.1.b Suppose the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Assumption 2: ϕ_k is independent of $\{w_s, s \geq k+1\}$ for $k = 0, 1, \ldots$

Assumption 3: $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ.

Let $f(w)$ be the probability density function for $\{w_k\}$. Then the MLE $\hat{\theta}_n$ maximizes

$$ln\ M_n(\theta) = \sum_{k=0}^{n-1} ln\ f(y_{k+1} - \phi_k^T \theta)$$

as shown in 2.2.1.a above. So, under Assumption 4, the MLE maximizes

$$ln\ M_n(\theta) = -\frac{1}{\sigma^2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=0}^{n-1} (y_{k+1} - \phi_k^T \theta)^2.$$

Since the first term on the right does not depend on θ, minimizing $ln\ M_n(\theta)$ above is to minimize $\sum_{k=0}^{n-1} (y_{k+1} - \phi_k^T \theta)^2$. Therefore, the MLE coincides with the LSE.

2.2.1.c Consider the ARMAX model of Equation (2) above, which is restated below.

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + \varepsilon_k + \sum_{i=1}^{p} c_i \varepsilon_{k-i}$$

Suppose the following assumption holds.

Assumption 5: $b_1 = \cdots = b_p = 0.$

Under Assumption 5, we have the ARMA model,

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} c_i \varepsilon_{k-i},$$

Let

$$\theta^o := (-a_1, \ldots, -a_p, c_1, \ldots, c_p)^T \in R^{2p},$$

$$w_{k+1} := \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, \varepsilon_k, \ldots, \varepsilon_{k+1-p})^T.$$

Then, the ARMAX model can be written as Equation (3), which is

$$y_{k+1} = \phi_k^T \theta^o + w_{k+1}.$$

Suppose the following assumption holds.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Let $\{\hat{h}_k z^k\}$ and $\{\hat{g}_k z^k\}$ be the power series defined in Equation (34) (See 3.2.5.b below for details). Since $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ (i.e., Assumption 3 holds), the MLE maximizes

$$M_n(\theta) = q^{\theta}_{1|0}(y_1 \mid \phi_0) q^{\theta}_{2|1}(y_2 \mid \phi_0, \phi_1, y_1) \cdots q^{\theta}_{n|n-1}(y_n \mid \phi_0, \ldots, \phi_{n-1}, y_1, \ldots, y_{n-1}).$$

Now,

$$\begin{aligned} q^{\theta}_{k|k-1}(y_k \mid \phi_s, y_s, s \leq k-1) &= q^{\theta}_{k|k-1}(\phi^T_{k-1}\, \theta^o + w_k \mid \phi_s, y_s, s \leq k-1) \\ &= q^{\theta}_{k|k-1}((y_{k-1}, \ldots, y_{k-p}, \varepsilon_{k-1}, \ldots, \varepsilon_{k-p})^T\, \theta^o + \varepsilon_k \\ &\qquad \mid \phi_s = (y_s, \ldots, y_{s+1-p}, \varepsilon_s, \ldots, \varepsilon_{s+1-p})^T, y_s, s \leq k-1) \\ &= q^{\theta}_{k|k-1}(\varepsilon_k \mid \phi_s, y_s, s \leq k-1) \end{aligned}$$

Since, as shown in Equation (30) in 3.2.5.b,

$$\varepsilon_k = \sum_{j=0}^{k} \hat{g}_{k-j} y_j.$$

Therefore, the MLE maximizes

$$M_n = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum_{k=1}^{n} (\sum_{j=0}^{k} \hat{g}_{k-j} y_j)^2]$$

or

$$ln\, M_n = -\frac{1}{\sigma^2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (\sum_{j=0}^{k} \hat{g}_{k-j} y_j))^2.$$

So, the MLE minimizes

$$\frac{\partial ln\, M_n}{\partial \hat{g}_i} = -\frac{1}{\sigma^2} \sum_{k=1}^{n} [y_{k-i}(\sum_{j=0}^{k} \hat{g}_{k-j} y_j)]$$

for all $0 \leq i \leq n$. Thus, the MLE coincides with the PEE that minimizes $V_n(\theta)$ given in Equation (31)

$$V_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} (y_k - \hat{y}_{k|k-1})^2 = \frac{1}{n} \sum_{k=1}^{n} (\sum_{j=0}^{k} \hat{g}_{n-j} y_j)^2$$

[1].

2.2.2 *The Least Squares Estimator (LSE)*

Suppose we have the data from the past $\{\phi_0, \phi_1, \ldots, \phi_{n-1}, y_1, \ldots, y_n\}$ where the ϕ_i and y_i are random variables such that

$$y_{k+1} \approx \phi^T_k\, \theta, \qquad k = 0, 1, 2, \ldots, n-1,$$

i.e., y_{k+1} is approximately a linear function of ϕ_k with parameter θ. The ϕ_i are vectors, and y_i are scalars. To estimate θ, we choose $\hat{\theta}$ to minimize

$$V_n(\theta) := \sum_{k=0}^{n-1} (y_{k+1} - \phi_k^T \theta)^2.$$

$$\partial V_n(\theta)/\partial\theta = 2 \sum_{k=0}^{n-1} (-\phi_k)(y_{k+1} - \phi_k^T \theta)$$

$$\partial V_n(\theta)/\partial\theta = 2 \sum_{k=0}^{n-1} (-\phi_k y_{k+1} + \phi_k \phi_k^T \theta)$$

Setting $\partial V_n(\theta)/\partial\theta = 0$, we get

$$-\sum_{k=0}^{n-1} \phi_k y_{k+1} + \sum_{k=0}^{n-1} \phi_k \phi_k^T \theta = 0$$

$$\sum_{k=0}^{n-1} \phi_k \phi_k^T \theta = \sum_{k=0}^{n-1} \phi_k y_{k+1}$$

So,

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1} \sum_{k=0}^{n-1} \phi_k y_{k+1} \tag{4}$$

assuming the inverse exists [1]. However, to compute the value of this estimate $\hat{\theta}_n$ at each step using the above formula (4) is obviously costly. Hence, we obtain $\hat{\theta}$ recursively as shown below.

Define $R_n := \sum_{k=0}^{n} \phi_k \phi_k^T$. Then the equation above becomes

$$\hat{\theta}_n = R_{n-1}^{-1} \sum_{k=0}^{n-1} \phi_k y_{k+1} \tag{5}$$

or

$$R_{n-1} \hat{\theta}_n = \sum_{k=0}^{n-1} \phi_k y_{k+1} \ . \tag{6}$$

So,

$$\begin{aligned} \hat{\theta}_{n+1} &= R_n^{-1} \sum_{k=0}^{n} \phi_k y_{k+1} \\ &= R_n^{-1} \sum_{k=0}^{n-1} \phi_k y_{k+1} + R_n^{-1} \phi_n y_{n+1} \\ &= R_n^{-1} R_{n-1} \hat{\theta}_n + R_n^{-1} \phi_n y_{n+1} && \text{from (6)} \\ &= R_n^{-1} (R_n - \phi_n \phi_n^T) \hat{\theta}_n + R_n^{-1} \phi_n y_{n+1} \end{aligned}$$

$$= \hat{\theta}_n - R_n^{-1}\phi_n\phi_n^T\hat{\theta}_n + R_n^{-1}\phi_n y_{n+1}$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n + R_n^{-1}\phi_n(y_{n+1} - \phi_n^T\hat{\theta}_n) \tag{7}$$

where

$$R_n = R_{n-1} + \phi_n\phi_n^T.$$

Now, we only need to evaluate $\phi_n(y_{n+1} - \phi_n^T\hat{\theta}_n)$ and $\phi_n\phi_n^T$ for each estimate at time n. Still, computing the inverse of the matrix R_n is costly. So, we introduce $P_n := R_n^{-1}$ as shown below. First, we apply the *matrix inversion lemma*

$$(N + GQH)^{-1} = N^{-1} - N^{-1}G(HN^{-1}G + Q^{-1})^{-1}HN^{-1},$$

to Equation (7).

[The matrix inversion lemma can be verified as follows.

$$(N + GQH)^{-1}(N + GQH) = [N^{-1} - N^{-1}G(HN^{-1}G + Q^{-1})^{-1}HN^{-1}](N + GQH)$$

$$I = (I + N^{-1}GQH) - N^{-1}G(HN^{-1}G + Q^{-1})^{-1}(H + HN^{-1}GQH)$$

$$N^{-1}GQH = N^{-1}G(HN^{-1}G + Q^{-1})^{-1}(H + HN^{-1}GQH)$$

$$I = (N^{-1}GQH)^{-1}N^{-1}G(HN^{-1}G + Q^{-1})^{-1}(H + HN^{-1}GQH)$$

$$I = (QH)^{-1}(HN^{-1}G + Q^{-1})^{-1}(H + HN^{-1}GQH)$$

$$I = (HN^{-1}GQH + Q^{-1}QH)^{-1}(H + HN^{-1}GQH)$$

$$I = (HN^{-1}GQH + H)^{-1}(HN^{-1}GQH + H)$$]

By setting $N = R_{n-1}$, $G = \phi_n$, $Q = 1$, and $H = \phi_n^T$, we get

$$P_n = P_{n-1} - P_{n-1}\phi_n(\phi_n^TP_{n-1}\phi_n + 1)^{-1}\phi_n^TP_{n-1}$$

$$P_n = P_{n-1} - \frac{P_{n-1}\phi_n\phi_n^TP_{n-1}}{1 + \phi_n^TP_{n-1}\phi_n}.$$

$$P_n = \frac{P_{n-1} + P_{n-1}\phi_n^TP_{n-1}\phi_n - P_{n-1}\phi_n\phi_n^TP_{n-1}}{1 + \phi_n^TP_{n-1}\phi_n}.$$

$$P_n = \frac{P_{n-1}}{1 + \phi_n^TP_{n-1}\phi_n}. \tag{8}$$

Then

$$\hat{\theta}_{n+1} = \hat{\theta}_n + P_n\phi_n(y_{n+1} - \phi_n^T\hat{\theta}_n) \tag{9}$$

or

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{P_{n-1}\phi_n}{1 + \phi_n^T P_{n-1}\phi_n}(y_{n+1} - \phi_n^T\hat{\theta}_n) \tag{10}$$

where we can recursively obtain P_n using Equation (8). Now, we no longer need to compute the inverse of R_n for each step.

2.2.3. Extended Least Squares Estimator (ELS)

Consider the ARMAX model of Equation (2) above, which is restated below.

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + \varepsilon_k + \sum_{i=1}^{p} c_i \varepsilon_{k-i}$$

or

$$\begin{aligned} &y_k + a_1 y_{k-1} + \cdots + a_p y_{k-p} \\ &\qquad = b_1 u_{k-1} + \cdots + b_p u_{k-p} + \cdots + \varepsilon_k + c_1\varepsilon_{k-1} + \cdots + c_k\varepsilon_{k-p} \end{aligned}$$

Let $\phi_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p}, \varepsilon_k, \ldots, \varepsilon_{k+1-p})^T$, $\theta := (-a_1, \ldots, a_p, b_1, \ldots, b_p, c_1, \ldots, c_p)^T$, and $w_k := \varepsilon_k$. Then we can rewrite the model as Equation (3),

$$y_k = \phi_{k-1}^T \theta + w_k$$

(or $y_{k+1} = \phi_k^T \theta + w_{k+1}$). This model looks the same as the model used for the LSE above, and the LSE is given in Equations (8) and (9), which are

$$P_n = \frac{P_{n-1}}{1 + \phi_n^T P_{n-1}\phi_n} ,$$

and

$$\hat{\theta}_{n+1} = \hat{\theta}_n + P_n\phi_n(y_{n+1} - \phi_n^T\hat{\theta}_n)$$

where $P_n = (\sum_{k=0}^{n} \phi_n\phi_n^T)^{-1}$.

However, since we cannot observe w_k, ϕ_n which contains w_k's is not available. So, w_n is estimated through the parameter estimate $\hat{\theta}_n$ as follows. Solving for w_n of Equation (3), we get

$$w_n = y_n - \phi_{n-1}^T\theta.$$

Now, since $w_n = \varepsilon_n$, define $\hat{\varepsilon}_n$ as below.

$$\hat{\varepsilon}_n = y_n - \hat{\phi}_{n-1}^T\hat{\theta}_n$$

where we define $\hat{\phi}_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p}, \hat{\varepsilon}_k, \ldots, \hat{\varepsilon}_{k+1-p})^T$. Now, Equation (7) can be modified as below.

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \hat{P}_n\hat{\phi}_n(y_{n+1} - \hat{\phi}_n^T\,\hat{\theta}_n) \tag{11}$$

and where $\hat{P}_n := (\sum_{k=0}^{n} \hat{\phi}_k\hat{\phi}_k^T)^{-1}$ and is obtained recursively by

$$\hat{P}_n = \frac{\hat{P}_{n-1}}{1 + \hat{\phi}_n^T\hat{P}_{n-1}\hat{\phi}_n} \,.$$

We could obtain an algorithm that is slightly simpler to implement if we let $\hat{\varepsilon}_n := y_n - \hat{\phi}_{n-1}\,\hat{\theta}_{n-1}$ so that we have $\hat{\theta}_{n+1} = \hat{\theta}_n + \hat{P}_n\hat{\phi}_{n+1}\hat{\varepsilon}_{n+1}$. However, the convergence properties change [1].

2.2.4. Weighted Least Squares Estimator (WLS)

Consider the ARMAX model of Equation (2) above, which is restated below.

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + \varepsilon_k + \sum_{i=1}^{p} c_i \varepsilon_{k-i}$$

where y_k, $\varepsilon_k \in R^l$, $u_k \in R^m$, and a_i, b_i, $c_i \in R$. Let (Ω, A, P) be a probability space and F_s be the σ-algebra generated by events occurring up to time s. Assume that the noise $\varepsilon = (\varepsilon_p)$ is a martingale difference sequence with

$$\sup_{s \geq 0} E\,[\;\|\varepsilon_{s+1}\|^2 \mid F_s] < \sigma^2 \;\text{ a.s.}$$

where σ^2 is deterministic. Let y_k be scalars (i.e., $l = 1$). Assume that the initial state $\phi_0^T = (y_0^p, u_0^{p+1}, \varepsilon_0^p)$ is F_0 measurable and, for $p \geq 0$,

$$\phi_k^T = (y_k^p, u_k^p, \varepsilon_k^p),$$

where $y_k^p = (y_k, \ldots, y_{k+1-p})$, $u_k^p = (u_k, \ldots, u_{k+1-p})$ and where $\varepsilon_k^p = (\varepsilon_k, \ldots, \varepsilon_{k+1-p})$. Let $\theta^T = (-a_1, \ldots, -a_p, b_1, \ldots, b_p, c_1, \ldots, c_p)^T$. Then we can rewrite the Equation (3) above as

$$y_{k+1} = \phi_k^T\,\theta + \varepsilon_{k+1}.$$

The noise ε is predicted by $\hat{\varepsilon}$ with $\hat{\varepsilon}_0^p = 0$ and, for $p \geq 0$,

$$\hat{\varepsilon}_{k+1} = y_{k+1} - \hat{\phi}_k^T\,\hat{\theta}_{k+1}$$

where

$$\hat{\phi}_k^T = (y_k^p, u_k^p, \hat{\varepsilon}_k^p).$$

So, we can also express the model as

$$y_{k+1} = \hat{\phi}_k^T \hat{\theta}_{k+1} + \hat{\varepsilon}_{k+1}.$$

Let $\{\alpha_k\}$ be a sequence of random variables adapted to F, positive, nonincreasing, and ≤ 1. To estimate θ, we choose $\hat{\theta}$ to minimize

$$V_n(\theta) := \sum_{k=0}^{n-1} \alpha_k(y_{k+1} - \phi_k^T \theta)^2.$$

$$(\text{or } \; V_n(\theta) := \sum_{k=0}^{n-1} \alpha_k(y_{k+1} - \hat{\phi}_k^T \theta)^2.)$$

$$\partial V_n(\theta)/\partial\theta = 2 \sum_{k=0}^{n-1} (-\alpha_k\phi_k)(y_{k+1} - \phi_k^T \theta)$$

$$\partial V_n(\theta)/\partial\theta = 2 \sum_{k=0}^{n-1} (-\alpha_k\phi_k y_{k+1} + \alpha_k\phi_k\phi_k^T \theta)$$

Setting $\partial V_n(\theta)/\partial\theta = 0$, we get

$$-\sum_{k=0}^{n-1} \alpha_k\phi_k y_{k+1} + \sum_{k=0}^{n-1} \alpha_k\phi_k\phi_k^T \theta = 0$$

$$\sum_{k=0}^{n-1} \alpha_k\phi_k\phi_k^T\theta = \sum_{k=0}^{n-1} \alpha_k\phi_k y_{k+1}$$

So,

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \alpha_k\phi_k\phi_k^T)^{-1} \sum_{k=0}^{n-1} \alpha_k\phi_k y_{k+1} \tag{12}$$

assuming the inverse exists [1]. Also, we can obtain $\hat{\theta}$ recursively as shown below.

Define $S_n := \sum_{k=0}^{n} \alpha_k\phi_k\phi_k^T$. Then the equation above becomes

$$\hat{\theta}_n = S_{n-1}^{-1} \sum_{k=0}^{n-1} \alpha_k\phi_k y_{k+1}$$

or

$$S_{n-1}\hat{\theta}_n = \sum_{k=0}^{n-1} \alpha_k\phi_k y_{k+1}.$$

So,

$$\begin{aligned}\hat{\theta}_{n+1} &= S_n^{-1} \sum_{k=0}^{n} \alpha_k\phi_k y_{k+1} \\ &= S_n^{-1} \sum_{k=0}^{n-1} \alpha_k\phi_k y_{k+1} + S_n^{-1}\alpha_n\phi_n y_{n+1} \\ &= S_n^{-1}S_{n-1}\hat{\theta}_n + S_n^{-1}\alpha_n\phi_n y_{n+1}\end{aligned}$$

$$= S_n^{-1}(S_n - \alpha_n\phi_n\phi_n^T)\hat{\theta}_n + S_n^{-1}\alpha_n\phi_n y_{n+1}$$

$$= \hat{\theta}_n - S_n^{-1}\alpha_n\phi_n\phi_n^T\,\hat{\theta}_n + S_n^{-1}\alpha_n\phi_n\,y_{n+1}$$

$$= \hat{\theta}_n + S_n^{-1}\alpha_n\phi_n(-\phi_n^T\,\hat{\theta}_n + y_{n+1})$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n + S_n^{-1}\phi_n\alpha_n(y_{n+1} - \phi_n^T\,\hat{\theta}_n) \tag{13}$$

and

$$S_n = S_{n-1} + \phi_n\alpha_n\phi_n^T. \tag{14}$$

However, finding the inverse of S_n for each n is costly. So, we introduce $U_n := S_n^{-1}$ as shown below. First, we apply the *matrix inversion lemma*

$$(N + GQH)^{-1} = N^{-1} - N^{-1}G(HN^{-1}G + Q^{-1})^{-1}HN^{-1},$$

to Equation (7). By setting $N = S_{n-1}$, $G = \phi_n$, $Q = \alpha_n$, and $H = \phi_n^T$, we get

$$U_n = U_{n-1} - U_{n-1}\phi_n(\phi_n^T U_{n-1}\phi_n + 1/\alpha_n)^{-1}\phi_n^T U_{n-1}$$

$$U_n = U_{n-1} - \frac{U_{n-1}\phi_n\phi_n^T U_{n-1}}{1/\alpha_n + \phi_n^T U_{n-1}\phi_n}.$$

$$U_n = \frac{(1/\alpha_n)U_{n-1} + U_{n-1}\phi_n^T U_{n-1}\phi_n - U_{n-1}\phi_n\phi_n^T U_{n-1}}{1/\alpha_n + \phi_n^T U_{n-1}\phi_n}.$$

$$U_n = \frac{U_{n-1}}{1 + \phi_n^T U_{n-1}\phi_n\alpha_n}. \tag{15}$$

Then

$$\hat{\theta}_{n+1} = \hat{\theta}_n + U_n\phi_n\alpha_n(y_{n+1} - \phi_n^T\hat{\theta}_n) \tag{16}$$

or

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{U_{n-1}\phi_n}{1/\alpha_n + \phi_n^T U_{n-1}\phi_n}(y_{n+1} - \phi_n^T\hat{\theta}_n) \tag{17}$$

where we can recursively obtain U_n using Equation (15) [2].

2.2.5 The Prediction Error Estimator (PEE)

Again, consider the model given in Equation (3), which is

$$y_{k+1} = \phi_k^T\,\theta + w_{k+1}.$$

We can predict y_{k+1} for the above model, given the past $\{y_s,\ \phi_s,\ s \le k\}$ by

$$\hat{y}_{k+1|k}(\theta) := E\,(y_{k+1} \mid y_s,\ \phi_s,\ s \leq k).$$

Since this prediction depends on θ, the predictor is a function of θ. We try to find an estimator which will make the prediction errors, $y_{k+1} - \hat{y}_{k+1|k}(\theta)$ smaller. The prediction error estimator is the estimator $\hat{\theta}_n$ which will minimize $L_n(\theta)$ defined as follows [1]:

$$L_n(\theta) := \frac{1}{2}\sum_{k=0}^{n-1} (y_{k+1} - \hat{y}_{k+1|k}(\theta))^2.$$

The PEE is a good estimate, but it is very hard to implement recursively. However, an approximation of the PEE is possible using the recursive prediction error method (RPEM) as follows.

The Recursive Prediction Error Method (RPEM):

By means of a Taylor expansion of $L_n(\theta)$ around $\hat{\theta}_{n-1}$, we obtain

$$L_n(\theta) = L_n(\hat{\theta}_{n-1}) + L_n'(\hat{\theta}_{n-1})(\theta - \hat{\theta}_{n-1}) + \frac{1}{2}\,(\theta - \hat{\theta}_{n-1})^T L_n''(\hat{\theta}_{n-1})(\theta - \hat{\theta}_{n-1}) + \mathrm{o}(|\theta - \hat{\theta}_{n-1}|^2),$$

where L_n' is the gradient and L_n'' is the matrix of second partial derivatives with respect to θ, and $\mathrm{o}(x)$ denotes a function such that $\mathrm{o}(x)/|x| \to 0$ as $|x| \to 0$. Let

$$\frac{\partial L_n(\theta)}{\partial \theta} = L_n'(\hat{\theta}_{n-1}) + [L_n''(\hat{\theta}_{n-1})\,(\theta - \hat{\theta}_{n-1})]^T + \mathrm{o}(|\theta - \hat{\theta}_{n-1}|^T) = 0$$

Then

$$L_n''(\hat{\theta}_{n-1})\,\theta = L_n''(\hat{\theta}_{n-1})\,\hat{\theta}_{n-1} - [L_n'(\hat{\theta}_{n-1})]^T + \mathrm{o}(|\theta - \hat{\theta}_{n-1}|).$$

So, plugging $\hat{\theta}_n$ to this equation, we have

$$\hat{\theta}_n = \hat{\theta}_{n-1} - [L_n''(\hat{\theta}_{n-1})]^{-1}\,[L_n'(\hat{\theta}_{n-1})]^T + \mathrm{o}(|\hat{\theta}_n - \hat{\theta}_{n-1}|). \tag{18}$$

Now, we need to approximate $L_n''(\hat{\theta}_{n-1})$ and $L_n'(\hat{\theta}_{n-1})$. Define

$$\zeta_k(\theta) := y_k - \hat{y}_{k|k-1}(\theta),$$

and

$$\psi_k(\theta) := [-\zeta_k'(\theta)]^T = [\hat{y}_{k|k-1}'(\theta)]^T, \tag{19}$$

where $\zeta_k'(\theta)$ is the gradient of $\zeta_k(\theta)$. Then, we have

$$[L_n'(\theta)]^T = [\frac{d}{d\theta}\ \frac{1}{2}\sum_{k=0}^{n-1} (\zeta_{k+1}(\theta))^2]^T$$

$$= \sum_{k=0}^{n-1} [\zeta'_{k+1}(\theta)]^T \zeta_{k+1}(\theta)$$

$$= -\sum_{k=0}^{n-1} \psi_{k+1}(\theta)\ \zeta_{k+1}(\theta)$$

$$= -\sum_{k=0}^{n-2} \psi_{k+1}(\theta)\ \zeta_{k+1}(\theta) - \psi_n(\theta)\ \zeta_n(\theta)$$

$$= [L'_{n-1}(\theta)]^T - \psi_n(\theta)\ \zeta_n(\theta) \tag{20}$$

By differentiating both sides again,

$$L''_n(\theta) = L''_{n-1}(\theta) + \psi_n(\theta)\psi_n^T(\theta) + \zeta''_n(\theta)\zeta_n(\theta) \tag{21}$$

where $\zeta''_n(\theta)$ is the second derivative matrix of $\zeta_n(\theta)$ with respect to θ. To approximate $L'_n(\theta)$ and $L''_n(\theta)$, suppose the following assumptions hold.

Assumption a: $\hat{\theta}_k$ is in a small neighborhood of $\hat{\theta}_{k+1}$.

Assumption b: $\zeta''_k(\hat{\theta}_{k-1})\zeta_k(\hat{\theta}_{k-1}) \approx 0$.

Assumption c: $\hat{\theta}_k$ is the optimal estimate at the time k.

Then, under Assumption b, Equation (21) becomes

$$L''_n(\theta) \approx L''_{n-1}(\theta) + \psi_n(\theta)\psi_n^T(\theta).$$

Under Assumption a, $L''_k(\hat{\theta}_k) \approx L''_k(\hat{\theta}_{k-1})$. So, R_n can be defined as an approximation of $L''_n(\theta)$ so that

$$R_n = R_{n-1} + \psi_n(\theta)\psi_n^T(\theta).$$

On the other hand, under Assumption c, $L'_k(\hat{\theta}_k) \approx 0$. So, by plugging $\hat{\theta}_{n-1}$ into Equation (20), we get

$$[L'_n(\hat{\theta}_{n-1})]^T \approx -\psi_n(\hat{\theta}_{n-1})\ \zeta_n(\hat{\theta}_{n-1}),$$

Therefore, by substituting $L''_n(\hat{\theta}_{n-1})$ and $L'_n(\hat{\theta}_{n-1})$ in Equation (18), and by setting $o(|\theta - \hat{\theta}_{n-1}|) = 0$ under Assumption a, we get

$$\hat{\theta}_n = \hat{\theta}_{n-1} + R_n^{-1}\ \psi_n(\hat{\theta}_{n-1})\ \zeta_n(\hat{\theta}_{n-1}),$$

Define $P_n := R_{\mathrm{n}}^{-1}$, then we can obtain

$$\hat{\theta}_n = \hat{\theta}_{n-1} + P_n \, \psi_n(\hat{\theta}_{n-1}) \, \zeta_n(\hat{\theta}_{n-1}), \tag{22}$$

where

$$P_n = P_{n-1} - \frac{P_{n-1}\psi_n\psi_n^T P_{n-1}}{1 + \psi_n^T P_{n-1}\psi_n} = \frac{P_{n-1}}{1 + \psi_n^T P_{n-1}\psi_n} \tag{23}$$

through the same procedure used get Equations (9) and (10) for the LSE [2]. Now, we need to find the way to recursively evaluate $\psi_{n-1}(\hat{\theta}_{n-1})$ and $\zeta_n(\hat{\theta}_{n-1})$.

Consider the ARMAX model,

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + \varepsilon_k + \sum_{i=1}^{p} c_i \varepsilon_{k-i},$$

or

$$\begin{aligned} y_k = -a_1\, y_{k-1} - \cdots - a_p\, y_{k-p} + b_1\, u_{k-1} + \cdots + b_p\, u_{k-p} \\ + \varepsilon_k + c_1\, \varepsilon_{k-1} + \cdots + c_p\, \varepsilon_{k-p}. \end{aligned}$$

for $k = 0, 1, 2, \ldots$

Let

$$\theta^{\circ} := (-a_1, \ldots, -a_p, b_1, \ldots, b_p, c_1, \ldots, c_p)^T,$$

$$w_{k+1} := \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p}, \varepsilon_k, \ldots, \varepsilon_{k+1-p})^T.$$

Then for the ARMAX system above,

$$\hat{y}_{k|k-1}(\theta) = \phi_{k-1}^T \, \theta = y_k - \varepsilon_k,$$

and

$$\varepsilon_k = y_k - \hat{y}_{k|k-1}(\theta) = \zeta_k(\theta).$$

Combining these two equations, we can write

$$\hat{y}_{k|k-1}(\theta) = (y_{k-1}, \ldots, y_{k-p}, u_{k-1}, \ldots, u_{k-p}, \zeta_{k-1}(\theta), \ldots, \zeta_{k-p}(\theta))\theta.$$

By the definition of $\psi_k(\theta)$ given in Equation (19), $\psi_k^T(\theta) = -\zeta_k'(\theta)$. So, if we differentiate both sides with respect to θ,

$$\psi_k^T(\theta) = (0, \ldots, 0, \zeta_{k-1}'(\theta), \ldots, \zeta_{k-p}'(\theta))(-a_1, \ldots, -a_p, b_1, \ldots, b_p, c_1, \ldots, c_p)^T + \phi_{k-1}^T$$

$$\psi_k^T(\theta) = c_1\zeta_{k-1}'(\theta) + \cdots + c_p\zeta_{k-p}'(\theta) + \phi_{k-1}^T$$

$$\psi_k^T(\theta) = -c_1\psi_{k-1}^T(\theta) - \cdots - c_p\psi_{k-p}^T(\theta) + \phi_{k-1}^T$$

So, we set

$$\varphi_k := -\hat{c}_1(k-1)\varphi_{k-1} - \cdots - \hat{c}_p(k-1)\varphi_{k-p} + \hat{\phi}_{k-1} \tag{24}$$

and

$$\hat{\zeta}_k(\theta) := y_k - \hat{\phi}_{k-1}^T\theta, \tag{25}$$

where $\hat{\phi}_{k-1} := (y_{k-1}, \ldots, y_{k-p}, u_{k-1}, \ldots, u_{k-p}, \hat{\zeta}_{k-1}(\hat{\theta}_{k-1}), \ldots, \hat{\zeta}_{k-p}(\hat{\theta}_{k-p}))^T$; $\varphi_k(\theta)$ and $\hat{\zeta}_k(\theta)$ approximate $\psi_k(\theta)$ and ζ_k, respectively; and where $(\hat{c}_1(k), \ldots, \hat{c}_p(k))$ is the last p components of $\hat{\theta}_k$. Thus, by Equations (22) and (23), the PEE can be approximated using Equations (24) and (25), together with (26) and (27) below.

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \hat{P}_n \, \varphi_n \, \hat{\zeta}_n(\hat{\theta}_{n-1}), \tag{26}$$

where $\hat{R}_n := \sum_{k=0}^{n} \varphi_k\varphi_k^T$ and $\hat{P}_n := \hat{R}_n^{-1}$. $\hat{P}_n$ can be recursively obtained by

$$\hat{P}_n = \hat{P}_{n-1} - \frac{\hat{P}_{n-1}\varphi_n\varphi_n^T\hat{P}_{n-1}}{1 + \varphi_n^T\hat{P}_{n-1}\varphi_n} = \frac{\hat{P}_{n-1}}{1 + \varphi_n^T\hat{P}_{n-1}\varphi_n} \tag{27}$$

[1].

2.2.5.a Suppose the following assumption holds.

Assumption 6: $E\,(w_{k+1} \mid y_s, \phi_s, s \leq k) = 0$.

Then the PEE is just the LSE as shown below.

$$\begin{aligned}
\hat{y}_{k+1|k}(\theta) &= E\,(y_{k+1} \mid y_s, \phi_s, s \leq k) \\
&= E\,(\phi_k^T\,\theta + w_{k+1} \mid y_s, \phi_s, s \leq k) \\
&= E\,(\phi_k^T\,\theta \mid y_s, \phi_s, s \leq k) + E\,(w_{k+1} \mid y_s, \phi_s, s \leq k) \\
&= E\,(\phi_k^T\,\theta \mid y_s, \phi_s, s \leq k) + 0 \\
&= \phi_k^T\,\theta
\end{aligned}$$

Then

$$L_n(\theta) = \sum_{k=0}^{n-1} (y_{k+1} - \phi_k^T\,\theta)^2 = V_n(\theta)$$

where $V_n(\theta)$ is as defined for the LSE in 2.2.2 above [1].

3. Comparing The Quality of Estimators

There are several measures to compare the quality of the estimators as follows. Consider the ARMAX model,

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} b_i u_{k-i} + \varepsilon_k + \sum_{i=1}^{p} c_i \varepsilon_{k-i},$$

or

$$y_k = -a_1\, y_{k-1} - \cdots - a_p\, y_{k-p} + b_1\, u_{k-1} + \cdots + b_p\, u_{k-p} \\ + \varepsilon_k + c_1\, \varepsilon_{k-1} + \cdots + c_p\, \varepsilon_{k-p}.$$

for $k = 0, 1, 2, \ldots$

3.1. Unbiasedness and Variances

Assume that there is no prior distribution over the set of model parameters Θ. Let $\tau:(y, u) \to \tau(y, u) \in R^p$ be any estimator, where $\theta \in R^p$. τ is **unbiased** if

$$E\,(\tau \mid u, \theta) = \int \tau(y, u)\, p(y \mid u, \theta)\, dy = \theta, \quad \text{for all } \theta \in \Theta.$$

Let $y^n = (y_1, y_2, \ldots, y_n)$ and $u^n = (u_1, u_2, \ldots, u_n)$. For each n, depending on y^n and u^n, $\tau_n(y^n, u^n)$ is the estimator. Then τ_n is **asymptotically unbiased** if

$$\lim_{n\to\infty} E\,(\tau_n \mid u^n, \theta) = \theta, \quad \text{for all } \theta \in \Theta.$$

3.1.1 Unbiasedness and Variances of The MLE

3.1.1.a Suppose the following assumptions hold.

Assumption 2: ϕ_k is independent of $\{w_s, s \geq k+1\}$ for $k = 0, 1, \ldots$

Assumption 3: $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Then, as shown in 2.2.1.b, the MLE coincides with the LSE.

3.1.1.b Suppose the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Assumption 5: $b_1 = \cdots = b_p = 0$.

Then, as shown in 2.2.1.c, the MLE coincides with the PEE.

3.1.2 Unbiasedness and Variances of The LSE

Let

$$\theta^{\mathrm{o}} := (-a_1, \ldots, -a_p, b_1, \ldots, b_p)^T \in \Theta = R^{2p},$$

$$w_{k+1} := c_1\, \varepsilon_k + \cdots + c_p\, \varepsilon_{k+1-p} + \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p})^T.$$

Then, the ARMAX model can be written as Equation (3), which is

$$y_{k+1} = \phi_k^T\, \theta^{\mathrm{o}} + w_{k+1}.$$

In 2.2 above, $\tau_n(y^n, \phi^n) = \hat{\theta}$ has been obtained in Equation (5) which is

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1} \sum_{k=0}^{n-1} \phi_k y_{k+1}\ .$$

Substituting y_{k+1} by Equation (3) gives

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1} \sum_{k=0}^{n-1} \phi_k (\phi_k^T\, \theta^{\mathrm{o}} + w_{k+1})$$

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1} (\sum_{k=0}^{n-1} \phi_k \phi_k^T)\, \theta^{\mathrm{o}} + (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1}\, \phi_k w_{k+1}$$

$$\hat{\theta}_n = \theta^{\mathrm{o}} + (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1} \sum\, \phi_k w_{k+1}$$

or

$$\hat{\theta}_n = \theta^{\mathrm{o}} + P_{n-1} \sum_{k=0}^{n-1} \phi_k w_{k+1} \tag{28}$$

where $P_n = (\sum_{k=0}^{n} \phi_k \phi_k^T)^{-1}$ as defined in 2.2.2.

unbiasedness:

3.1.2.a Suppose the following assumption holds.

Assumption 7: $E\,(w_{k+1} \mid \phi_s,\ s \leq \infty) = 0,\ \ k = 0, 1, \ldots$

Then,

$$
\begin{aligned}
E\,\hat{\theta}_n &= E\,(E\,(\theta^o + P_{n-1}\sum_{k=0}^{n-1}\phi_k w_{k+1} \mid \phi_s,\ s \le n-1)) \\
&= E\,(\theta^o + E\,(P_{n-1}\sum_{k=0}^{n-1}\phi_k w_{k+1} \mid \phi_s,\ s \le n-1)) \\
&= \theta^o + E\,(P_{n-1}\sum_{k=0}^{n-1}\phi_k\, E\,(w_{k+1} \mid \phi_s,\ s \le n-1)) \\
&= \theta^o + 0 \\
&= \theta^o
\end{aligned}
$$

Hence, the LSE is unbiased [1].

3.1.2.b Suppose the following assumptions hold.

Assumption 5: $b_1 = \cdots = b_p = 0$.

Assumption 9: $a_1 = \cdots = a_p = 0$.

Assumption 10: $E\,(\varepsilon_k\varepsilon_j \mid \phi_s,\ s \le \infty) = \sigma^2\delta_{kj}$ where σ^2 is unknown and δ_{kj} is the **Kronecker delta**[1].

Assumption 11: $c_1 \neq 0,\ c_2 = \cdots = c_p = 0$, and $\theta^o = 0$.

Under the assumptions above, we have

$$y_{k+1} = c_1\,\varepsilon_k + \varepsilon_{k+1},$$

or, we can write

$$y_{k+1} = \phi_k\theta^o + w_{k+1},$$

by defining $w_k := c_1\,\varepsilon_k + \varepsilon_{k+1}$ and $\phi_k = y_k \in R$. Then

$$
\begin{aligned}
\hat{\theta}_n &= (\sum_{k=0}^{n-1}\phi_k\phi_k^T)^{-1}\sum_{k=0}^{n-1}\phi_k y_{k+1} && \text{from (4)} \\
&= (\sum_{k=0}^{n-1} y_k^2)^{-1}\sum_{k=0}^{n-1} y_k y_{k+1}
\end{aligned}
$$

[1] δ_{kj} is 1 if $k = j$ and 0 if $k \neq j$.

$$= [\frac{1}{n-1} \sum_{k=0}^{n-1} (c_1 \varepsilon_{k-1} + \varepsilon_k)^2]^{-1} [\frac{1}{n-1} \sum_{k=0}^{n-1} (c_1 \varepsilon_{k-1} + \varepsilon_k)(c_1 \varepsilon_k + \varepsilon_{k+1})]$$

$$= [\frac{1}{n-1} \sum_{k=0}^{n-1} (c_1^2 \varepsilon_{k-1}^2 + 2c_1 \varepsilon_{k-1}\varepsilon_k + \varepsilon_k^2)]^{-1}$$

$$\times [\frac{1}{n-1} \sum_{k=0}^{n-1} (c_1^2 \varepsilon_{k-1}\varepsilon_k + c_1 \varepsilon_k^2 + c_1 \varepsilon_{k-1}\varepsilon_{k+1} + \varepsilon_k w_{k+1})]$$

Now, consider $X_n := \sum_{k=0}^{n} \frac{\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}}{k}$. We can show that this is a martingale as below.

$$E\,(X_{n+1} \mid \varepsilon_s,\, s \le n) = E\,(\sum_{k=0}^{n} \frac{\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}}{k} \mid \varepsilon_s,\, s \le n) + E\,(\frac{\varepsilon_{n+1}\varepsilon_{n+1-m} - \sigma^2\delta_{m0}}{n+1})$$

$$= \sum_{k=0}^{n} \frac{\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}}{k} + 0$$

$$= X_n$$

Also,

$$E\,(|X_n|^2) \le E\,(\sum_{k=0}^{n} \frac{|\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}|^2}{k^2})$$

$$\le \sum_{k=0}^{n} \frac{E\,|\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}|^2}{k^2} = 0 < \infty,$$

since there is no negative terms. Then, by the martingale convergence theorem,

$$\lim_{n\to\infty} X_n = \sum_{k=0}^{\infty} \frac{\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}}{k} < \infty \quad \text{a.s.}$$

Then, by Kronecker's Lemma,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} (\varepsilon_k\varepsilon_{k-m} - \sigma^2\delta_{m0}) = 0 \quad \text{a.s.}$$

or

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \varepsilon_k\varepsilon_{k-m} = \sigma^2\delta_{m0} \quad \text{a.s.}$$

So,

$$\lim_{n\to\infty} E\,\hat{\theta}_n = (c_1^2\sigma^2 + 2c_1\delta_{10} + \sigma^2)^{-1}(c_1^2\delta_{10} + c_1\sigma^2 + c_1\delta_{20} + \delta_{10})$$

$$= \frac{c_1\sigma^2}{c_1^2\sigma^2 + \sigma^2} = \frac{c_1}{c_1^2 + 1} \ne 0 = \theta^{o}$$

Hence, the LSE is asymptotically biased.

In general, the LSE of parameters $(-a_1, \ldots, -a_p, b_1, \ldots, b_p)$ will be asymptotically biased unless $c_1 = \cdots = c_p = 0$ [1].

variances:

3.1.2.c Suppose the following assumption holds.

Assumption 12: $E\ (w_k w_j \mid \phi_s,\ s \leq \infty) = \sigma^2 \delta_{kj}$ where σ^2 is unknown and δ_{kj} is the Kronecker delta.

Then,

$$\begin{aligned} E\ (\hat{\theta}_n - \theta^o)(\hat{\theta}_n - \theta^o)^T &= P_{n-1}\ E\ (\sum_{k=0}^{n-1} \phi_k \sigma^2 \phi_k^T)\ P_{n-1} \\ &= P_{n-1}\ (\sigma^2\ P_{n-1}^{-1})\ P_{n-1} \\ &= \sigma^2\ P_{n-1} \qquad (29) \end{aligned}$$

3.1.3 Unbiasedness and Variances of The ELS

The ELS is obtained by Equation (11)

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \hat{P}_n \phi_n (y_{n+1} - \hat{\phi}_n^T\, \hat{\theta}_n),$$

where $\hat{\phi}_k := (y_k,\ \ldots,\ y_{k+1-p},\ u_k,\ \ldots,\ u_{k+1-p},\ \hat{\varepsilon}_k,\ \ldots,\ \hat{\varepsilon}_{k+1-p})^T$, $\hat{w}_n = \hat{\varepsilon}_n = y_n - \hat{\phi}_{n-1}^T \hat{\theta}_n$ and $\hat{P}_n := (\sum_{k=0}^{n} \hat{\phi}_k \hat{\phi}_k^T)^{-1}$.

unbiasedness:

3.1.3.a Suppose the following assumption holds.

Assumption 12: $E\ (\hat{w}_{k+1} \mid \hat{\phi}_s,\ s \leq \infty) = 0,\ \ k = 0,\ 1,\ \ldots$

Then,

$$\begin{aligned} E\ \hat{\theta}_n &= E\ (E\ (\theta^o + \hat{P}_{n-1} \sum_{k=0}^{n-1} \hat{\phi}_k \hat{w}_{k+1} \mid \hat{\phi}_s,\ s \leq n-1)) \\ &= E\ (\theta^o + E\ (\hat{R}_{n-1} \sum_{k=0}^{n-1} \hat{\phi}_k \hat{w}_{k+1} \mid \hat{\phi}_s,\ s \leq n-1)) \\ &= \theta^o + E\ (\hat{P}_{n-1} \sum_{k=0}^{n-1} \hat{\phi}_k\, E\ (\hat{w}_{k+1} \mid \hat{\phi}_s,\ s \leq n-1)) \\ &= \theta^o + 0 \\ &= \theta^o \end{aligned}$$

Hence, the ELS is unbiased.

3.1.4 Unbiasedness and Variances of The WLS

Let

$$\theta^o := (-a_1, \ldots, -a_p, b_1, \ldots, b_p)^T \in \Theta = R^{2p},$$

$$w_{k+1} := c_1\, \varepsilon_k + \cdots + c_p\, \varepsilon_{k+1-p} + \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p})^T.$$

Then, the ARMAX model can be written as Equation (3), which is

$$y_{k+1} = \phi_k^T\, \theta^o + w_{k+1}.$$

unbiasedness:

Inserting Equation (3) to Equation (11) below, we get

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \alpha_k \phi_k \phi_k^T)^{-1} \sum_{k=0}^{n-1} \alpha_k \phi_k\, y_{k+1}$$

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \alpha_k \phi_k \phi_k^T)^{-1} [\sum_{k=0}^{n-1} \alpha_k \phi_k (\phi_k^T\, \theta^o + w_{k+1})]$$

$$\hat{\theta}_n = \theta^o + (\sum_{k=0}^{n-1} \alpha_k \phi_k \phi_k^T)^{-1} \sum_{k=0}^{n-1} \alpha_k \phi_k w_{k+1} \qquad (30)$$

So, the WLS is unbiased if ϕ_{k+1} and w_k are independent [2].

3.1.4.a Suppose the following assumption holds.

Assumption 7: $E\,(w_{k+1} \mid \phi_s,\, s \leq \infty) = 0,\;\; k = 0, 1, \ldots$

Then, the WLS is unbiased. This can be shown by the same procedure as used in 3.1.2.a.

variances:

3.1.4.b Suppose the following assumption holds.

Assumption 12: $E\,(w_k w_j \mid \phi_s,\, s \leq \infty) = \sigma^2 \delta_{kj}$ where σ^2 is unknown and δ_{kj} is the Kronecker delta.

Then,

$$\begin{aligned} E\,(\hat{\theta}_n - \theta^o)(\hat{\theta}_n - \theta^o)^T &= U_{n-1}\, E\,(\sum_{k=0}^{n-1} \phi_k \sigma^2 \alpha_k^2 \phi_k^T)\, U_{n-1} \\ &= U_{n-1}\sigma^2 (\sum_{k=0}^{n-1} \phi_k \alpha_k^2 \phi_k^T)\, U_{n-1} \\ &\leq U_{n-1}\sigma^2 (\sum_{k=0}^{n-1} \alpha_k \phi_k \phi_k^T)\, U_{n-1} \\ &\leq \sigma^2 U_{n-1} \end{aligned} \tag{31}$$

since $0 < \alpha_k \leq 1$ by definition.

3.1.5 Unbiasedness and Variances of The PEE

3.1.5.a Suppose the following assumption holds.

Assumption 14: $E\,(w_{k+1} \mid F_k) = 0,\ k = 0, 1, \ldots,$
where F_k is the σ-algebra generated by $\{y_s, \phi_s, s \leq k\}$ ($\Rightarrow$ *Assumption 6*).

Then, as shown in 2.2.5, the PEE is same as the LSE.

3.2 Consistency

Obviously, we need to make sure that the estimator converges after an infinite number of iterations. If it does, the speed of convergence for each estimators should be compared.

For the same estimator $\hat{\theta}_n$ defined in 3.1. above, $\hat{\theta}_n$ is **asymptotically consistent** if $\tau_n(y^n, u^n)$ converges to θ in probability. Also, $\hat{\theta}_n$ is **strongly consistent** if $\tau_n(y^n, u^n)$ converges to θ with probability 1 for all θ, where $y^n := (y_1, \ldots, y_n)$ and $u^n := (u_1, \ldots, u_n)$.

Let

$$\theta^o := (-a_1, \ldots, -a_p, b_1, \ldots, b_p)^T \in \Theta = R^{2p},$$

$$w_{k+1} := c_1\, \varepsilon_k + \cdots + c_p\, \varepsilon_{k+1-p} + \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p})^T.$$

Then, the ARMAX model can be written as Equation (3), which is

$$y_{k+1} = \phi_k^T \theta^o + w_{k+1}.$$

3.2.1 Consistency of The MLE

3.2.1.a Suppose the following assumptions hold.

Assumption 2: ϕ_k is independent of $\{w_s,\ s \geq k+1\}$ for $k = 0, 1, \ldots$

Assumption 3: $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Then, as shown in 2.2.1.b, the MLE coincides with the LSE.

3.2.1.b Suppose the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Assumption 5: $b_1 = \cdots = b_p = 0$.

Then, as shown in 2.2.1.c, the MLE coincides with the PEE.

3.2.2 Consistency of The LSE

asymptotical consistency:

First, we note that

$$\begin{aligned}(\hat{\theta}_n - \theta^{\circ})(\hat{\theta}_n - \theta^{\circ})^T &= (P_{n-1} \sum_{k=0}^{n-1} \phi_k w_{k+1})(P_{n-1} \sum_{k=0}^{n-1} \phi_k w_{k+1})^T \qquad \text{from (28)} \\ &= P_{n-1} (\sum_{k=0}^{n-1} \phi_k w_{k+1})(\sum_{k=0}^{n-1} \phi_k w_{k+1})^T P_{n-1} \\ &= P_{n-1} (\phi_0 w_1 + \cdots + \phi_{n-1} w_n)(\phi_0 w_1 + \cdots + \phi_{n-1} w_n)^T P_{n-1} \\ &= P_{n-1} (\sum_{k=0}^{n-1} \sum_{j=0}^{n-1} \phi_k w_{k+1} w_{j+1} \phi_j^T) P_{n-1}\end{aligned}$$

3.2.2.a Suppose that the following assumptions hold.

Assumption 13: $E\,(w_k w_j \mid \phi_s,\ s \leq \infty) = \sigma^2 \delta_{kj}$ where σ^2 is unknown and δ_{kj} is the Kronecker delta.

Assumption 15: $\lim_{n \to \infty} E\,(Tr\,P_{n-1}) = 0.$

Under Assumption 13, Equation (29),

$$E\,(\hat{\theta}_n - \theta^{\text{o}})(\hat{\theta}_n - \theta^{\text{o}})^T = \sigma^2\,P_{n-1},$$

holds as shown in 3.2.1.c. Since

$$\|S\|^2 = \sum_{i=1}^{r} s_i^2 = Tr(SS^T) \qquad (32)$$

for any vector $S = (s_1 \cdots s_r)$,

$$\begin{aligned} E\,(\|\hat{\theta}_n - \theta^{\text{o}}\|^2) &= E\,(Tr\,[(\hat{\theta}_n - \theta^{\text{o}})(\hat{\theta}_n - \theta^{\text{o}})^T]) \\ &= Tr\,[\,E(\hat{\theta}_n - \theta^{\text{o}})(\hat{\theta}_n - \theta^{\text{o}})^T] \\ &= Tr\,[\,E\,(\sigma^2\ \ P_{n-1})] \\ &= \sigma^2\,E\,(Tr\,P_{n-1}) \end{aligned}$$

By Chebyshev's inequality, for every $\varepsilon > 0$,

$$\frac{E\,(\|\hat{\theta}_n - \theta^{\text{o}}\|^2)}{\varepsilon^2} \geq P\,(\|\hat{\theta}_n - \theta^{\text{o}}\| \geq \varepsilon)$$

$$\frac{\delta^2\,E\,(Tr\,P_{n-1})}{\varepsilon^2} \geq P\,(\|\hat{\theta}_n - \theta^{\text{o}}\| \geq \varepsilon)$$

Taking the limit of both sides of this inequality, under Assumption 15,

$$0 \geq \lim_{n \to \infty} P\,(\|\hat{\theta}_n - \theta^{\text{o}}\| \geq \varepsilon) \geq 0$$

for every $\varepsilon > 0$. So,

$$\lim_{n \to \infty} P\,(\|\hat{\theta}_n - \theta^{\text{o}}\| \geq \varepsilon) = 0$$

Hence, $\hat{\theta}_n$ is asymptotically consistent [1].

3.2.2.b Suppose that the following assumptions hold.

Assumption 9: $a_1 = \cdots = a_p = 0.$

Assumption 13: $E\,(w_k w_j \mid \phi_s,\, s \leq \infty) = \sigma^2 \delta_{kj}$ where σ^2 is unknown and δ_{kj} is the Kronecker delta.

Assumption 16: $\{u_k\}$ is deterministic.

Assumption 17: The input sequence $\{u_k\}$ is **persistently exciting** of order $p + 1$; i.e., there is a positive definite $(p + 1) \times (p + 1)$ matrix U and an integer N such that for all $n \geq N$,

$$\frac{1}{n} P_{n-1}^{-1} \geq U.$$

Under Assumption 7, we let $\phi_k := (u_k, \ldots, u_{k-p})^T$. Then, by Assumption 17,

$$n P_{n-1} \leq \frac{1}{U}$$

$$Tr (P_{n-1}) \leq \frac{1}{n \, Tr \, U}$$

Let $\varepsilon > 0$, then we can find some positive number $M \geq N$ such that $1/M \leq \varepsilon$. Then, since $Tr \, U > 0$, for any $n > M$,

$$Tr (P_{n-1}) \leq \frac{1}{n \, Tr \, U} \leq \varepsilon.$$

Then, $lim_{n \to \infty} E \, (Tr \, P_{n-1}) = 0$; i.e., Assumption 15 holds. Hence, as shown in 3.2.2.a, $\hat{\theta}_n$ is asymptotically consistent [1].

strong consistency:

Let F_k be the σ-algebra generated by the past $\{y_s, u_s, w_s, s \leq k\}$. Since $w_k = y_k - \phi_{k-1}^T \, \theta^o$, F_k is also the σ-algebra generated by $\{y_s, \phi_s, s \leq k\}$. Also, let

$$R_n = \sum_{k=0}^{n} \phi_k \phi_k^T.$$

3.2.2.c Suppose the following assumptions hold. Let F_k be the σ-algebra generated by $\{y_s, \phi_s, s \leq k\}$.

Assumption 8: $c_1 = \cdots = c_p = 0$.

Assumption 14: $E \, (w_{k+1} \mid F_k) = 0, \; k = 0, 1, \ldots$ ($\Rightarrow$ *Assumption 6*)

Assumption 18: $E \, (w_{k+1}^2 \mid F_k) \leq \sigma^2$.

Assumption 19: $\lim_{n \to \infty} \lambda_{min} R_n = +\infty$.

Assumption 20: $\dfrac{R_n}{Tr \, R_n} \geq \varepsilon I$ for all large n and some $\varepsilon > 0$.

Under Assumption 7, we have the general ARX model,

$$y_k = -a_1 \, y_{k-1} - \cdots - a_p \, y_{k-p} + b_1 \, u_{k-1} + \cdots + b_p \, u_{k-p} + w_k.$$

The LSE is obtained in Equation (28),

$$\hat{\theta}_n = \theta^o + P_{n-1} \sum_{k=0}^{n-1} \phi_k w_{k+1} = \theta^o + R_{n-1}^{-1} \sum_{k=0}^{n-1} \phi_k w_{k+1}$$

where $P_n := R_n^{-1}$. Consider the term $\sum_{k=0}^{n-1} \phi_k w_{k+1}$ in the above equation. Let ϕ_k^i be the ith component of the vector ϕ_k. For each i, define $S_0^i := 1$, $S_{-1}^i := 1$, and

$$S_n^i := 1 + \sum_{k=0}^{n} (\phi_k^i)^2, \quad n = 1, 2, \ldots$$

So, S_n^i is adapted to F_n. Now, define $z_0^i := 0$, and

$$z_n^i = \sum_{k=0}^{n-1} \frac{\phi_k^i \, w_{k+1}}{S_k^i}, \quad n = 1, 2, \ldots$$

for $i = 1, \ldots, 2p + 2$. Then z_n^i is also adapted to F_n and

$$E\,(z_{n+1}^i \mid F_n) = E\,(z_n^i + \frac{\phi_n^i \, w_{n+1}}{S_n^i} \mid F_n) = z_n^i + \frac{\phi_n^i}{S_n^i} E\,(w_{n+1} \mid F_n) = z_n^i$$

Thus, $\{z_n^i, F_n\}$ is a martingale. Then, by the martingale convergence theorem,

$$\lim_{n\to\infty} z_n^i = \lim_{n\to\infty} \sum_{k=0}^{n-1} \frac{\phi_k^i \, w_{k+1}}{S_k^i} \text{ exists and is finite a.s.} \tag{33}$$

On the other hand, Assumption 19 implies

$$S_{n-1}^i := 1 + \sum_{k=0}^{n-1} (\phi_k^i)^2 = +\infty. \tag{34}$$

Therefore, by Equations (33) and (34), we can apply *Kronecker's Lemma* to get

$$\lim_{n\to\infty} \frac{1}{S_{n-1}^i} \sum_{k=0}^{n-1} \phi_k^i \, w_{k+1} = 0, \quad i = 1, \ldots, 2p{+}2 \text{ a.s.}$$

Now,

$$\begin{aligned} S_{n-1}^i &:= 1 + \sum_{k=0}^{n-1} (\phi_k^i)^2 \le 1 + \sum_{k=0}^{n-1} \sum_{j=1}^{2p+2} (\phi_k^j)^2 = 1 + Tr \sum_{k=0}^{n-1} \phi_k \phi_k^T \\ &= 1 + Tr\,(R_{n-1}), \quad i = 1, \ldots, 2p + 2 \end{aligned}$$

So,

$$0 = \lim_{n\to\infty} \frac{1}{S_{n-1}^i} \sum_{k=0}^{n-1} \phi_k^i \, w_{k+1} \ge \lim_{n\to\infty} \frac{1}{1 + Tr\,(R_{n-1})} \sum_{k=0}^{n-1} \phi_k^i \, w_{k+1} \ge 0$$

or

$$\lim_{n\to\infty} \frac{1}{1 + Tr\,(R_{n-1})} \sum_{k=0}^{n-1} \phi_k^i \, w_{k+1} = 0$$

But, Equation (28) can be written as

$$\hat{\theta}_n = \theta^o + \left[\frac{1}{Tr\,(R_{n-1})} R_{n-1} \right]^{-1} \times \left[\frac{1}{Tr\,(R_{n-1})} \sum_{k=0}^{n-1} \phi_k w_{k+1} \right]$$

Also, under Assumption 20, there is an integer $N \geq 0$ such that for any $n \geq N$,

$$\left[\frac{R_{n-1}}{Tr\,R_{n-1}} \right]^{-1} \leq \frac{1}{\varepsilon} I.$$

Hence, $\lim_{n\to\infty} \hat{\theta}_n = \theta^o$ a.s.

3.2.3 Consistency of The ELS

strong consistency:

3.2.3.a Define the polynomials

$$A(z) := 1 + a_1 z + \cdots + a_p z^p,$$

$$C(z) := 1 + c_1 z + \cdots + c_p z^p,$$

and let $\phi_k = (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p}, \varepsilon_k, .., \varepsilon_{k+1-p})$. Suppose the following assumptions hold.

Assumption 21: $E\,(\varepsilon_{k+1} \mid \varepsilon_s, s \leq k) = 0, \;\; k = 0, 1, \ldots$

Assumption 22: $E\,(\varepsilon_{k+1}^2 \mid \varepsilon_s, s \leq k) = \sigma^2.$

Assumption 23: $E\,(\varepsilon_{k+1}^4 \mid \varepsilon_s, s \leq k) = \delta.$

Assumption 24: All the roots of the polynomials $A(z)$ and $C(z)$ are strictly outside the closed unit disc.

Assumption 25: $\mathrm{Re}[\frac{1}{C\,(e^{i\varepsilon})} - \frac{1}{2}] \geq 0$ for all ε,

where Re means real part and $i := \sqrt{-1}.$

Assumption 26: The input sequence $\{u_k\}$ is adapted to F_k.

Assumption 27: $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi_k \phi_k^T = P$, where $P > 0$.

Then $\lim\limits_{n\to\infty} \hat{\theta}_n = \theta^o$ a.s.

3.2.4 Consistency of The WLS

asymptotical consistency:

As shown in Equation (30) in 3.1.4,

$$\hat{\theta}_n = \theta^o + (\frac{1}{n-1}\sum_{k=0}^{n-1}\alpha_k\phi_k\phi_k^T)^{-1}\frac{1}{n-1}\sum_{k=0}^{n-1}\alpha_k\phi_k w_{k+1}.$$

Under weak conditions, the sum $(1/n-1)\ \sum_{k=0}^{n-1}\alpha_k\phi_k w_{k+1}$ will converge to its expected value as n approaches infinity, according to the law of large numbers. This expected value depends on the correlation between the disturbance term w_k and the vector ϕ_k. It is zero only if w_k and ϕ_k are uncorrelated [2].

strong consistency:

Let F_k be the σ-algebra generated by the past $\{y_s, u_s, w_s, s \le k\}$. Since $w_k = y_k - \phi_{k-1}^T\theta^o$, F_k is also the σ-algebra generated by $\{y_s, \phi_s, s \le k\}$. Also, let

$$S_n = \sum_{k=0}^{n}\phi_k\phi_k^T.$$

3.2.2.c Suppose the following assumptions hold. Let F_k be the σ-algebra generated by $\{y_s, \phi_s, s \le k\}$.

Assumption 8: $c_1 = \cdots = c_p = 0.$

Assumption 14: $E(w_{k+1} \mid F_k) = 0,\ k = 0, 1, \ldots$ (⇒ *Assumption 6*)

Assumption 18: $E(w_{k+1}^2 \mid F_k) \le \sigma^2.$

Assumption 28: $\lim\limits_{n\to\infty}\lambda_{min} S_n = +\infty.$

Assumption 29: $\dfrac{S_n}{Tr\ S_n} \ge \varepsilon I$ for all large n and some $\varepsilon > 0$.

For each i, define $\tilde{S}_0^i := 1$, $\tilde{S}_{-1}^i := 1$, and

$$\tilde{S}_n^i := 1 + \sum_{k=0}^{n}\alpha_k(\phi_k^i)^2,\ n = 1, 2, \ldots$$

Also, define $\tilde{z}_0^i := 0$, and

$$\tilde{z}^i_n = \sum_{k=0}^{n-1} \frac{\phi^i_k \, w_{k+1}}{S^i_k}, \quad n = 1, 2, \ldots$$

for $i = 1, \ldots, 2p + 2$. Then, by a similar procedure as shown in 3.2.2.c for the strong consistency of the LSE, it can be shown that $lim\ \hat{\theta}_n = \theta^o$ a.s.

3.2.5 Consistency of The PEE

3.2.5.a Suppose the following assumption holds.

Assumption 14: $E\ (w_{k+1} \mid F_k) = 0,\ \ k = 0, 1, \ldots$ (⇒ *Assumption 6*)

Then, as shown in 2.2.5, PEE is same as LSE.

strong consistency:

3.2.5.b Define the polynomials

$$A(z) := 1 + a_1 z + \cdots + a_p z^p,$$

$$C(z) := 1 + c_1 z + \cdots + c_p z^p,$$

and suppose the following assumptions hold.

Assumption 5: $b_1 = \cdots = b_p = 0.$

Assumption 21: $E\ (\varepsilon_{k+1} \mid \varepsilon_s,\ s \le k) = 0,\ \ k = 0, 1, \ldots$

Assumption 22: $E\ (\varepsilon^2_{k+1} \mid \varepsilon_s,\ s \le k) = \sigma^2.$

Assumption 23: $E\ (\varepsilon^4_{k+1} \mid \varepsilon_s,\ s \le k) = \delta.$

Assumption 24: All the roots of the polynomials $A(z)$ and $C(z)$ are strictly outside the closed unit disc.

Assumption 30: $\varepsilon_{-p} = \cdots = \varepsilon_0 = 0$, and $y_{-p} = \cdots = y_0 = 0$

Under Assumption 5, we have the ARMA model,

$$y_k + \sum_{i=1}^{p} a_i y_{k-i} = \sum_{i=1}^{p} c_i \varepsilon_{k-i},$$

or

$$y_k = -a_1\, y_{k-1} - \cdots - a_p\, y_{k-p} + \varepsilon_k + c_1\, \varepsilon_{k-1} + \cdots + c_p\, \varepsilon_{k-p}. \tag{35}$$

Let

$$\theta^{\circ} := (-a_1, \ldots, -a_p, c_1, \ldots, c_p)^T \in R^{2p},$$

$$w_{k+1} := \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, \varepsilon_k, \ldots, \varepsilon_{k+1-p})^T.$$

Then, the ARMAX model can be written as Equation (3), which is

$$y_{k+1} = \phi_k^T \theta^{\circ} + w_{k+1}.$$

Using the shift operator q, Equation (35) above can be written as

$$y_k = -a_1 q^{-1} y_k - \cdots - a_p q^{-p} y_k + \varepsilon_k + c_1 q^{-1} \varepsilon_k + \cdots + c_p q^{-p} \varepsilon_k.$$

So,

$$A(q^{-1}) y_k = C(q^{-1}) \varepsilon_k. \tag{36}$$

where

$$A(q^{-1}) := 1 + a_1 q^{-1} + \cdots + a_p q^{-p},$$

$$C(q^{-1}) := 1 + c_1 q^{-1} + \cdots + c_p q^{-p}.$$

Let $\{h_k z^k\}$ and $\{g_k z^k\}$ be the power series such that

$$\sum_{k=0}^{\infty} h_k z^k := \frac{C(z)}{A(z)} \quad \text{and} \quad \sum_{k=0}^{\infty} g_k z^k := \frac{A(z)}{C(z)}. \tag{37}$$

However, $A(z)$ and $C(z)$ are unknown. So, we define the polynomials

$$\hat{A}(z) := 1 + \hat{a}_1 z + \cdots + \hat{a}_{\hat{p}} z^{\hat{p}},$$

$$\hat{C}(z) := 1 + \hat{c}_1 z + \cdots + \hat{c}_{\hat{p}} z^{\hat{p}},$$

where

$$\hat{\theta}_n := (-\hat{a}_1, \ldots, -\hat{a}_{\hat{p}}, \hat{c}_1, \ldots, \hat{c}_{\hat{p}})^T \in \Theta \subset R^{2\hat{p}}.$$

Also, for $\theta \in \Theta$, let $\{\hat{h}_k z^k\}$ and $\{\hat{g}_k z^k\}$ be the power series

$$\sum_{k=0}^{\infty} \hat{h}_k z^k := \frac{\hat{C}(z)}{\hat{A}(z)} \quad \text{and} \quad \sum_{k=0}^{\infty} \hat{g}_k z^k := \frac{\hat{A}(z)}{\hat{C}(z)}. \tag{38}$$

Note that $\hat{h}_0 = \hat{g}_0 = 1$.

Suppose the following assumption holds.

Assumption 31: Θ is a compact set such that for every θ in Θ, all roots of the polynomials $\hat{A}(z)$ and $\hat{C}(z)$ are strictly outside the unit disc.

Now, from Equation (36),

$$\varepsilon_n = \frac{\hat{A}(z)}{\hat{C}(z)} y_n = \sum_{j=0}^{n} \hat{g}_j q^j y_n = \sum_{j=0}^{n} \hat{g}_{n-j} y_j, \tag{39}$$

and

$$y_n = \frac{\hat{C}(z)}{\hat{A}(z)} \varepsilon_n = \sum_{j=0}^{n} \hat{h}_j q^j \varepsilon_n = \sum_{j=0}^{n} \hat{h}_{n-j} \varepsilon_j.$$

So,

$$\begin{aligned}
\hat{y}_{n|n-1}(\theta) &= E\,(y_n \mid y_s,\, s \leq n-1) \\
&= E\,(\sum_{j=0}^{n} \hat{h}_{n-j} \varepsilon_j \mid y_s,\, s \leq n-1) \\
&= E\,(\sum_{j=0}^{n-1} \hat{h}_{n-j} \varepsilon_j + \hat{h}_0 \varepsilon_n \mid y_s,\, s \leq n-1) \\
&= E\,(\sum_{j=0}^{n-1} \hat{h}_{n-j} \varepsilon_j \mid y_s,\, s \leq n-1) + E\,(\varepsilon_n \mid y_s,\, s \leq n-1) \\
&= E\,(\sum_{j=0}^{n} \hat{h}_{n-j} \varepsilon_j - \hat{h}_0 \varepsilon_n \mid y_s,\, s \leq n-1) \\
&= E\,(y_n - \varepsilon_n \mid y_s,\, s \leq n-1) \\
&= E\,(y_n - \sum_{j=0}^{n} \hat{g}_{n-j} y_j \mid y_s,\, s \leq n-1) \\
&= E\,(y_n - \sum_{j=0}^{n-1} \hat{g}_{n-j} y_j - \hat{g}_0 y_n \mid y_s,\, s \leq n-1) \\
&= -\sum_{j=0}^{n-1} \hat{g}_{n-j} y_j
\end{aligned}$$

Hence

$$y_k - \hat{y}_{k|k-1} = y_k + \sum_{j=0}^{k-1} \hat{g}_{n-j} y_j = \sum_{j=0}^{k} \hat{g}_{n-j} y_j.$$

So, the PEE $\hat{\theta}_n$ is $\theta \in \Theta$ that will minimize $L_n(\theta)$ below.

$$L_n(\theta) := \sum_{k=1}^{n} (y_k - \hat{y}_{k|k-1})^2$$

Let $V_n(\theta) := (1/n)L_n$. Then the PEE minimizes

$$V_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} (y_k - \hat{y}_{k|k-1})^2 = \frac{1}{n} \sum_{k=1}^{n} (\sum_{j=0}^{k} \hat{g}_{n-j}\, y_j)^2 \tag{40}$$

Now, in order to show the a.s. convergence of the estimate, we first show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} y_m y_{m-l} = \sigma^2 \sum_{k=0}^{\infty} h_k\, h_{k+l} \text{ a.s.}$$

Since $y_m = \sum_{j=0}^{m} h_j\, w_{m-j}$,

$$\begin{aligned} \frac{1}{n} \sum_{m=0}^{n} y_m\, y_{m-l} &= \frac{1}{n} \sum_{m=0}^{n} (\sum_{j=0}^{m} h_j\, \varepsilon_{m-j})(\sum_{k=0}^{m-l} h_k\, \varepsilon_{m-l-k}) \\ &= \frac{1}{n} \sum_{m=0}^{n} (\sum_{j=0}^{m} \sum_{k=0}^{m-l} h_j\, h_k\, \varepsilon_{m-j}\, \varepsilon_{m-l-k}) \end{aligned}$$

Since

$$0 \le k \le m - l \le n - l \quad \text{and} \quad 0 \le j \le m \le n, \tag{41}$$

we can rewrite this as

$$\frac{1}{n} \sum_{m=0}^{n} y_m\, y_{m-l} = \frac{1}{n} \sum_{k=0}^{n-l} \sum_{j=0}^{n} h_j\, h_k \sum_{m=0}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}$$

However, as shown in Equation (41), $j \le m$ and $k + l \le m$. So,

$$\frac{1}{n} \sum_{m=0}^{n} y_m\, y_{m-l} = \sum_{k=0}^{n-l} \sum_{j=0}^{n} h_j\, h_k \frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}. \tag{42}$$

Now, for $|z| \le 1$, $C(z) \neq 0$ and $A(z) \neq 0$. So, by the definition given in Equation (37), and by Assumption 24, $\sum_{k=0}^{\infty} h_k\, z^k$ converges for $|z| \le 1$. Let

$$\gamma := \lim_{j\to\infty} \sup \sqrt[j]{|h_j|}\ .$$

Then

$$|z| \le 1 < R = \frac{1}{\gamma}.$$

So,

$$0 < \gamma < 1,$$

where R is the radius of convergence. Now, there exists J such that $j \geq J$ implies

$$\sqrt[j]{|h_j|} < \gamma$$

or

$$|h_j| < \gamma^{i}.$$

Let $Q = \max\{|h_j| \mid j = 1, \ldots, J-1\}$. Then we can find a constant $\alpha > 0$ such that

$$|h_j| \leq Q = \alpha\, \gamma^j \quad \text{and} \quad \gamma^j \leq \alpha\, \gamma^j. \tag{43}$$

Then $|h_j| \leq \alpha\, \gamma^j$ for all j. Also, by the Schwarz inequality,

$$\begin{aligned} \left|\frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}\right|^2 &\leq \frac{1}{n^2} \sum_{m=1}^{n} |\varepsilon_{m-j}|^2 \sum_{m=1}^{n} |\varepsilon_{m-l-k}|^2 \\ &\leq \frac{1}{n^2} \left(\sum_{m=1}^{n} |\varepsilon_m|^2\right)^2 \end{aligned}$$

So,

$$\left|\frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}\right| \leq \frac{1}{n} \sum_{m=1}^{n} \varepsilon_m^2 \leq M \tag{44}$$

for some random constant M. So,

$$\begin{aligned} \left|\frac{1}{n} \sum_{m=0}^{n} y_m\, y_{m-l}\right|^2 &\leq \sum_{k=0}^{n-l} \left| \sum_{j=0}^{n} h_j\, h_k\, \frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}\right|^2 \\ &\leq \sum_{k=0}^{n-l} \left[\left| \sum_{j=0}^{n} h_j\, h_k\right|^2 \left|\frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}\right|^2\right] \\ &\leq \sum_{k=0}^{n-l} \left[\sum_{j=0}^{n} |h_j|^2\, |h_k|^2 \left|\frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j}\, \varepsilon_{m-l-k}\right|^2\right] \end{aligned}$$

By Equations (43) and (44),

$$\left|\frac{1}{n} \sum_{m=0}^{n} y_m\, y_{m-l}\right|^2 \leq \sum_{k=0}^{n-l} \sum_{j=0}^{n} |\alpha^2 \gamma^j\, \gamma^k|^2\, M.$$

So, there is a constant $\tilde{M}$ such that

$$\left|\frac{1}{n} \sum_{m=0}^{n} y_m\, y_{m-l}\right| \leq \tilde{M} \tag{45}$$

Now, consider $X_n := \sum_{k=0}^{n} \dfrac{\varepsilon_k\, \varepsilon_{k-m} - \sigma^2 \delta_{m0}}{k}$. Under Assumptions 21 and 22,

$$E\,(\varepsilon_{k+1}\varepsilon_{k+1-l} \mid \phi_s,\ s \leq k) = \sigma^2 \delta_{kj}$$

where δ_{kj} is the Kronecker delta. Therefore, as shown in 3.1.2.b,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \varepsilon_k \, \varepsilon_{k-m} = \sigma^2 \delta_{m0} \quad \text{a.s.} \tag{46}$$

Then, by Equations (42) and (45),

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n} y_m \, y_{m-l} = \lim_{n\to\infty} \sum_{k=0}^{n-l} \sum_{j=0}^{n} h_j \, h_k \, (\lim_{n\to\infty} \frac{1}{n} \sum_{m=\max(j,\, k+l)}^{n} \varepsilon_{m-j} \, \varepsilon_{m-l-k}) \quad \text{a.s.}$$

Then, by Equation (46), only the terms with $j = k + l$ are non-zero. Hence,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n} y_m \, y_{m-l} = \lim_{n\to\infty} \sigma^2 \sum_{k=0}^{n-l} h_k \, h_{k+l} = \sigma^2 \sum_{k=0}^{\infty} h_k \, h_{k+l} \quad \text{a.s.} \tag{47}$$

Now, rewriting Equation (40), we get

$$V_n(\theta) = \frac{1}{n} \sum_{k=0}^{n} (\sum_{j=0}^{k} \sum_{l=0}^{k} \hat{g}_j \, \hat{g}_l \, y_{k-j} \, y_{k-l})$$

Since, $0 \le j \le k \le n$, $0 \le l \le k \le n$, $k-0 \le k-j \le k-n$, and $k-0 \le k-l \le k-n$, by letting $m := \max\{k, j\}$, it can be rewritten

$$\begin{aligned} V_n(\theta) &= \frac{1}{n} \sum_{j=0}^{n} \sum_{k=0}^{n} \sum_{m=1}^{n} \hat{g}_j \, \hat{g}_k \, y_m \, y_{m-|j-k|} \\ &= \frac{1}{n} \sum_{j=0}^{n} \sum_{k=0}^{n} \hat{g}_j \, \hat{g}_k \sum_{m=1}^{n} y_m \, y_{m-|j-k|} . \end{aligned}$$

Now, under Assumption 29, $\sum_{k=0}^{\infty} \hat{g}_k \, z^k$ converges for $|z|$ for all $\theta \in \Theta$. So, by the same procedure which is used to obtain Equation (40), we can find constants β and $0 < \eta < 1$ such that $|\hat{g}_j| \le \beta\eta^j$ for all $\theta \in \Theta$. Then, by Equation (42) and Schwarz inequality,

$$|V_n(\theta)| \le \frac{1}{n} \sum_{j=0}^{n} \sum_{k=0}^{n} |\beta^2 \eta^{\mathrm{j}} \eta^{\mathrm{k}}|^2 \tilde{M}$$

Then, by Equation (44),

$$\begin{aligned} \lim_{n\to\infty} V_n(\theta) &= \lim_{n\to\infty} \sum_{j=0}^{n} \sum_{k=0}^{n} \hat{g}_j \, \hat{g}_k \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} y_m \, y_{m-|j-k|} . \\ &= \lim_{n\to\infty} \sum_{j=0}^{n} \sum_{k=0}^{n} \hat{g}_j \, \hat{g}_k \, (\alpha^2 \sum_{l=0}^{\infty} h_l \, h_{l+|j-k|}) \quad \text{a.s.} \end{aligned}$$

$$= \alpha^2 \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \hat{g}_j \hat{g}_k h_l h_{l+|j-k|} \text{ a.s.}$$

Hence, the PEE $\hat{\theta}_n$ converges to Θ a.s. [1].

The consistency of the RPEM:

As same as for the off-line PEE, the RPEM gives convergence under very general conditions [2].

3.3. Efficiency

The efficiency of the estimators can be compared using the following definition.

Definition: The **Fisher information matrix** is the $p \times p$ matrix $I(u,\theta)$ with elements

$$I_{ij}(u,\theta) := \int \frac{\partial ln\, p(y|u,\theta)}{\partial \theta_i} \frac{\partial ln\, p(y|u,\theta)}{\partial \theta_j} p(y|u,\theta)dy.$$

The Cramér-Rao Inequality

Let $\hat{\theta}$ be any unbiased estimator. Then

$$[I(u, \theta)]^{-1} \leq \text{Cov}(\hat{\theta} \mid u, \theta) \text{ for all } \theta \in \Theta.$$

The estimator θ is **efficient** if $[I(u, \theta)]^{-1} = \text{Cov}(\hat{\theta} \mid u, \theta)$. Now, consider the following estimators for the model given in Equation (3)

$$y_{k+1} = \phi_k^T \theta^o + w_{k+1}.$$

Let

$$\theta^o := (-a_1, \ldots, -a_p, b_1, \ldots, b_p)^T \in \Theta = R^{2p},$$

$$w_{k+1} := c_1 \varepsilon_k + \cdots + c_p \varepsilon_{k+1-p} + \varepsilon_{k+1},$$

and

$$\phi_k := (y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-p})^T.$$

Then, the ARMAX model can be written as Equation (3), which is

$$y_{k+1} = \phi_k^T \theta^o + w_{k+1}.$$

3.3.1 Efficiency of The MLE

3.3.1.a Suppose the following assumptions hold.

Assumption 2: ϕ_k is independent of $\{w_s, s \geq k+1\}$ for $k = 0, 1, \ldots$

Assumption 3: $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2

Then, as shown in 2.2.1.b, the MLE coincides with the LSE.

3.2.1.b Suppose the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Assumption 5: $b_1 = \cdots = b_p = 0$.

Then, as shown in 2.2.1.c, the MLE coincides with the PEE.

3.3.2 Efficiency of The LSE

3.3.2.a Suppose that the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2

Assumption 8: $c_1 = \cdots = c_p = 0$.

Assumption 9: $a_1 = \cdots = a_p = 0$.

Assumption 16: $\{u_k\}$ is deterministic.

Under Assumptions 8 and 9, $\phi_k := (u_k, \ldots, u_{k-p})^T$. Also, since $E\,(w_{k+1} \mid \phi_s, s \leq \infty) = 0$ under the assumptions above, $\hat{\theta}_n$ is unbiased (See **3.1.2.a** above), and $E\,(w_k w_j \mid \phi_s, s \leq \infty) = \sigma^2 \delta_{kj}$ where σ^2 is unknown and δ_{kj} is the Kronecker delta. Let $I\,(\theta^o)$ be the Fisher information matrix. Then by the Cramér-Rao Inequality stated above,

$$[I\,(\theta^o)]^{-1} \leq E\,(\hat{\theta}_n - \theta^o)(\hat{\theta}_n - \theta^o)^T$$

Also, from Equation (28) in 3.1.2., $E\,(\hat{\theta}_n - \theta^o)(\hat{\theta}_n - \theta^o)^T = \sigma^2\, P_{n-1}$. So,

$$[I\,(\theta^o)]^{-1} \leq \sigma^2\, P_{n-1}.$$

Now, we need to show that the equality holds. Since $\{w_k\} \sim N(0, \sigma^2)$,

$$p(y^n \mid \theta^o) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\,[\frac{-(y_k - \phi_{k-1}^T\,\theta^o)^2}{2\sigma^2}],$$

$$ln\, p(y^n \mid \theta^o) = -\frac{n}{2}\, ln(2\pi\sigma^2) - \sum_{k=1}^{n} \frac{(y_k - \phi_{k-1}^T\,\theta^o)^2}{2\sigma^2}$$

$$\frac{\partial ln\, p(y^n \mid \theta^o)}{\partial\theta_i^o} = \sum_{k=1}^{n} \frac{1}{\sigma^2}\,(y_k - \phi_{k-1}^T\,\theta^o)\, u_{k-i} = \sum_{k=1}^{n} \frac{w_k\, u_{k-i}}{\sigma^2}$$

$$\frac{\partial ln\, p(y^n \mid \theta^o)}{\partial\theta_i^o} = \sum_{k=0}^{n} \frac{w_{k+1}\, u_{k-i}}{\sigma^2}$$

So, the information matrix $I(\theta^o)$ is

$$\begin{aligned} I_{ij}(\theta^o) :&= \int \frac{\partial ln\, p(y^n \mid \theta^o)}{\partial\theta_i^o}\, \frac{\partial ln\, p(y^n \mid \theta^o)}{\partial\theta_j^o}\, p(y^n \mid \theta^o) dy^n \\ &= E^\theta\,(\sum_{k=0}^{n-1} \frac{w_{k+1}\, u_{k-i}}{\sigma^2} \sum_{l=0}^{n-1} \frac{w_{l+1}\, u_{l-j}}{\sigma^2}) \\ &= \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \sigma^2 u_{k-i}\, u_{k-j} \\ &= \frac{1}{\sigma^2} \sum_{k=0}^{n-1} u_{k-i}\, u_{k-j} \\ &= \frac{1}{\sigma^2}\, P_{n-1}^{-1} \end{aligned} \tag{48}$$

Therefore, $[I\,(\theta^o)]^{-1} = E\,(\hat\theta_n - \theta^o)(\hat\theta_n - \theta^o)^T$. Hence, $\hat\theta_n$ is efficient [1].

3.3.3 Efficiency of The ELS

3.3.3.a Suppose that the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2

Assumption 8: $c_1 = \cdots = c_p = 0$.

Assumption 9: $a_1 = \cdots = a_p = 0$.

Assumption 16: $\{u_k\}$ is deterministic.

Under Assumptions 8 and 9, $\phi_k := (u_k, \ldots, u_{k-p})^T = \hat{\phi}_k$. Hence, as shown in 3.3.2.a, $\hat{\theta}_n$ is efficient.

3.3.4 Efficiency of The WLS

3.3.4.a Suppose that the following assumptions hold.

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2

Assumption 8: $c_1 = \cdots = c_p = 0$.

Assumption 9: $a_1 = \cdots = a_p = 0$.

Assumption 16: $\{u_k\}$ is deterministic.

As shown in Equation (48) in 3.3.2.a,

$$I_{ij}(\theta^o) = \frac{1}{\sigma^2} P_{n-1}^{-1}.$$

Also, as shown in Equation (31) in 3.1.4.b, since $0 < \alpha_k \leq 1$,

$$E\,(\hat{\theta}_n - \theta^o)(\hat{\theta}_n - \theta^o)T \leq \sigma^2 U_{n-1} \leq \sigma^2 P_{n-1}$$

where $P_{n-1} = (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1}$ and $U_{n-1} = (\sum_{k=0}^{n-1} \alpha_k \phi_k \phi_k^T)^{-1}$. So, unless $\alpha_k = 1$ for all $k = 0, 1, \ldots$ (i.e., the WLS is just the LSE), $\hat{\theta}_n$ is not efficient.

3.3.5 Efficiency of The PEE

3.3.5.a Suppose the following assumption holds.

Assumption 14: $E\,(w_{k+1} \mid F_k) = 0$, $k = 0, 1, \ldots$ ($\Rightarrow$ *Assumption 6*)

Then, as shown in 2.2.4, PEE is same as LSE.

4. Examples

Example 1: Suppose that

$$y_{k+1} = -a\ u_k - c\varepsilon_k + \varepsilon_{k+1},\ \ k = 0,\ 1,\ \ldots$$

where y_k, $u_k \in R$; $\{\varepsilon_k\} \sim N(0,\sigma^2)$, $\{u_k\} \sim N(\mu,\delta^2)$ and are iid; $\varepsilon_0 = 0$; and $\{\varepsilon_k\}$ is independent of $\{u_k\}$. Set $a^o = 0$.

(a) Assume that *c* is known:

Let $\phi_k := u_k$, $\theta := -a$, and $w_{k+1} := -c\varepsilon_k + \varepsilon_{k+1}$ so that we have

$$y_{k+1} = \phi_k\theta + w_{k+1}.$$

The MLE: Since $\{u_k\}$ doesn't depend on θ, maximizing the likelihood function L is same as maximizing

$$p^\theta(y_s,\ s \leq n) = q^\theta_{1|0}(y_1 \mid u_0)q^\theta_{2|1}(y_2 \mid u_0,\ u_1,\ y_1) \cdots q^\theta_{n|n-1}(y_n \mid u_0,\ \ldots,\ u_{n-1},\ y_1,\ \ldots,\ y_{n-1})$$

Now, we can show that

$$\varepsilon_{k+1} = \sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\ u_j),$$

or

$$y_{k+1} = \theta\ u_k - c\sum_{j=0}^{k-1} c^{k-1-j}(y_{j+1} - \theta\ u_j) + \varepsilon_{k+1},$$

for $k \geq 1$. The proof can be done by induction. We know

$$\varepsilon_1 = y_1 - \theta\ u_0.$$

Assume that

$$\varepsilon_k = \sum_{j=0}^{k-1} c^{k-1-j}(y_{j+1} - \theta\ u_j).$$

Then

$$\begin{aligned} y_{k+1} &= \theta\ u_k + w_{k+1} \\ &= \theta\ u_k - c\varepsilon_k + \varepsilon_{k+1} \\ &= \theta\ u_k - c\sum_{j=0}^{k-1} c^{k-1-j}(y_{j+1} - \theta\ u_j) + \varepsilon_{k+1} \end{aligned}$$

$$\varepsilon_{k+1} = y_{k+1} - \theta\, u_k + c \sum_{j=0}^{k-1} c^{k-1-j}(y_{j+1} - \theta\, u_j)$$
$$= \sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\, u_j).$$

Thus, for $k \geq 1$,

$$q^{\theta}_{k+1|k}(y_{k+1} \mid u_0, \ldots, u_k, y_1, \ldots, y_k)$$
$$= P(\theta\, u_k - c \sum_{j=0}^{k-1} c^{k-1-j}(y_{j+1} - \theta\, u_j) + \varepsilon_{k+1} \mid u_0, \ldots, u_k, y_1, \ldots, y_k)$$
$$= P(\varepsilon_{k+1} \mid u_0, \ldots, u_k, y_1, \ldots, y_k),$$

and for $k = 0$,

$$P(y_1 \mid u_0) = P(\theta\, u_0 + \varepsilon_1) = P(\varepsilon_1).$$

Then, since $\{\varepsilon_k\} \sim N(0,\sigma^2)$,

$$p(\varepsilon_1, \ldots,\varepsilon_n) = \prod_{k=0}^{n-1} f(\varepsilon_{k+1}) = (2\pi\sigma^2)^{-n/2}\exp[-\frac{1}{2\sigma^2} \sum^{n-1} \varepsilon^2_{k+1}]$$

So, the MLE maximizes

$$M_n = (2\pi\sigma^2)^{-n/2}\exp[-\frac{1}{2\sigma^2} \sum_{k=0}^{n-1} (\sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\, u_j))^2]$$

$$ln\, M_n = -\frac{1}{\sigma^2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=0}^{n-1} (\sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\, u_j))^2$$

$$\frac{\partial ln\, M_n}{\partial\theta} = -\frac{1}{\sigma^2} \sum_{k=0}^{n-1} [-(\sum_{j=0}^{k} c^{k-j}\, u_j)(\sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\, u_j))] = 0$$

$$\sum_{k=0}^{n-1} [(\sum_{j=0}^{k} c^{k-j}\, u_j)(\sum_{j=0}^{k} c^{k-j}\, y_{j+1} - \sum_{j=0}^{k} c^{k-j}\, u_j\, \theta))] = 0$$

$$\sum_{k=0}^{n-1} [(\sum_{j=0}^{k} c^{k-j}\, u_j)(\sum_{j=0}^{k} c^{k-j}\, y_{j+1})] = \sum_{k=0}^{n-1} [(\sum_{j=0}^{k} c^{k-j}\, u_j)(\sum_{j=0}^{k} c^{k-j}\, u_j\, \theta)]$$

$$\hat{\theta}_n = \frac{\sum_{k=0}^{n-1} [(\sum_{j=0}^{k} c^{k-j}\, u_j)(\sum_{j=0}^{k} c^{k-j}\, y_{j+1})]}{\sum_{k=0}^{n-1} (\sum_{j=0}^{k} c^{k-j}\, u_j)^2}$$

or

$$\hat{\theta} = \frac{\sum_{k=0}^{n-1} [(u_k + \sum_{j=0}^{k-1} c^{k-j} u_j)(y_{k+1} + \sum_{j=0}^{k-1} c^{k-j} y_{j+1})]}{\sum_{k=0}^{n-1} (u_k + \sum_{j=0}^{k} c^{k-j} u_j)^2}$$

If $c = 0$ then this is same as the LSE. We can obtain the MLE recursively. Let

$$\tilde{\phi}_k = \sum_{j=0}^{k} c^{k-j} u_j, \quad \text{for } k = 0, 1, 2, \ldots$$

$$\tilde{y}_{k+1} = \sum_{j=0}^{k} c^{k-j} y_{j+1}, \quad \text{for } k = 0, 1, 2, \ldots$$

and

$$\tilde{R}_{\mathrm{n}} = \sum_{k=0}^{n} \tilde{\phi}_k^2 \quad \text{for } n = 0, 1, 2, \ldots$$

Then we get

$$\hat{\theta}_n = \tilde{R}_{n-1}^{-1} \sum_{k=0}^{n-1} \tilde{\phi}_k \tilde{y}_{k+1}$$

or

$$\tilde{R}_{n-1}\hat{\theta}_n = \sum_{k=0}^{n-1} \tilde{\phi}_k \tilde{y}_{k+1} \; .$$

Now, we can refer to the calculation done in 2.2.2. below, in which we started from Equation (5), which is

$$\hat{\theta}_n = R_{n-1}^{-1} \sum_{k=0}^{n-1} \phi_k y_{k+1}$$

where $R_n = \sum_{k=0}^{n} \phi_k \phi_k^T$, and obtained Equation (8),

$$\hat{\theta}_{n+1} = \hat{\theta}_n + P_n \phi_n (y_{n+1} - \phi_n^T \hat{\theta}_n),$$

where $P_n := R_n^{-1}$ and

$$P_n = \frac{P_{n-1}}{1 + \phi_n^T P_{n-1} \phi_n} \; . \qquad \text{from (7)}$$

Replacing y_{k+1} by $\tilde{y}_{k+1}$, P_n by $\tilde{P}_n$, ϕ_n by $\tilde{\phi}_n$, we have

$$\tilde{y}_{k+1} = c \sum_{j=0}^{k} c^{k-1-j} \tilde{y}_{j+1} + \tilde{y}_{k+1} = c\tilde{y}_k + \tilde{y}_{k+1},$$

$$\tilde{\phi}_k = c \sum_{j=0}^{k-1} c^{k-1-j}\, u_j + u_k = c\tilde{\phi}_{k-1} + u_k,$$

and

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \tilde{P}_n\tilde{\phi}_n(\tilde{y}_{n+1} - \tilde{\phi}_n^T\, \hat{\theta}_n),$$

where $\tilde{P}_n := \tilde{R}_n^{-1}$. P_n is obtained recursively by

$$\tilde{P}_n = \frac{\tilde{P}_{n-1}}{1 + \tilde{\phi}_n^2\tilde{P}_{n-1}} \ .$$

The LSE: Using Equation (4) in 2.2.2., we get

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} u_k^{\,2})^{-1} \sum_{k=0}^{n-1} u_k y_{k+1}.$$

Also, the recursive estimator is

$$\hat{\theta}_{n+1} = \hat{\theta}_n + P_n(y_{n+1} - \hat{\theta}_n u_n),$$

where

$$P_n = (\sum_{k=0}^{n} u_k^2)^{-1} = \frac{P_{n-1}}{1 + P_{n-1}u_n^2} \ ,$$

as shown in Equations (8) and (9). This estimate is biased and converges to $-c/(c^2 + 1)$ as shown in 3.1.2.b.

The PEE:

$$\begin{aligned}\hat{y}_{k+1|k}(\theta) &= E\,(y_{k+1} \mid y_s, u_s, s \le k)\\ &= E\,(\theta\, u_k + w_{k+1} \mid y_s, u_s, s \le k)\\ &= E\,(\theta\, u_k - c\varepsilon_k + \varepsilon_{k+1} \mid y_s, u_s, s \le k)\\ &= \theta\, u_k - cE\,(\varepsilon_k \mid y_s, u_s, s \le k) + E\,(\varepsilon_{k+1} \mid y_s, u_s, s \le k)\\ &= \theta\, u_k - cE\,(\varepsilon_k \mid y_s, u_s, s \le k)\end{aligned}$$

since $E\,(\varepsilon_{k+1} \mid y_s, u_s, s \le k) = E\,(\varepsilon_{k+1}) = 0$. Now, we have shown that

$$E\,(\varepsilon_k \mid y_s, u_s, s \le k) = \varepsilon_k = \sum_{j=0}^{k-1} c^{k-1-j}(y_{j+1} - \theta\, u_j).$$

So,

$$\hat{y}_{k+1|k}(\theta) = \theta\, u_k - \sum_{j=0}^{k-1} c^{k-j}(y_{j+1} - \theta\, u_j).$$

$$L_n(\theta) = \sum_{k=0}^{n-1} (y_{k+1} - \theta\, u_k - \sum_{j=0}^{k-1} c^{k-j}(y_{j+1} - \theta\, u_j))^2$$

$$L_n(\theta) = \sum_{k=0}^{n-1} (\sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\, u_j))^2$$

$$\frac{\partial L}{\partial \theta} = 2 \sum_{k=0}^{n-1} [- (\sum_{j=0}^{k} c^{k-j}\, u_j)(\sum_{j=0}^{k} c^{k-j}(y_{j+1} - \theta\, u_j))] = 0$$

But, the solution to the above equation is the same as the solution for the equation $\partial ln\, M_n / \partial \theta = 0$ for the MLE. So, the PEE is same as the MLE for this system.

Simulation: (See **Figure 1a.**)

The simulation has been done using the LSE and the MLE (= the PEE for this system) for a parameter a, with the true parameter values $a^o = 0$ and $c^o = 0.8$. The system is

$$y_{k+1} = -(0)u_k - 0.8\varepsilon_k + \varepsilon_{k+1},$$

where $\{\varepsilon_k\} \sim N(0,1)$, $\{u_k\} \sim N(1,1)$; or

$$y_{k+1} = \theta^o\, u_k + w_{k+1},$$

where $\theta^o = -a^o$ and $w_{k+1} = -0.8\varepsilon_k + \varepsilon_{k+1}$. The combinations of two different initial parameter values, $\hat{\theta}_0 = -0.5$ and 0; and two different initial values for P_0, $P_0 = 0.01$ and 1.0, were used.

In general, the MLE (or the PEE) seems to behave in a more stable manner, and to converge to the true parameter quicker than the LSE. This is not surprising since the value of c is in the algorithm for the MLE, while the LSE ignores it.

Also, the size of P_0 effects the speed of convergence, as will be discussed later using Example 2 [2].

(b) Assume *c* is unknown:

Let $\phi_k := (u_k, \varepsilon_k)^T$, $\theta := (-a, -c)^T$, and $w_{k+1} = \varepsilon_{k+1}$, $k = 0, 1, \ldots$ so that we have

$$y_{k+1} = \phi_k^T\, \theta + w_{k+1}.$$

The MLE: The following assumptions hold:

Assumption 4: $\{w_k\}$ is independently normally distributed with the mean 0 and the variance σ^2.

Assumption 2: ϕ_k is independent of $\{w_s, s \geq k+1\}$ for $k = 0, 1, \ldots$

Assumption 3: $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ.

So, as shown in 2.2.1.b, the MLE coincides with the LSE.

The ELS: Let $\hat{\varepsilon}_k := y_k - \hat{\phi}^T_{k-1}\,\hat{\theta}_k$ where $\hat{\phi}_k := (u_k, \hat{\varepsilon}_k)^T$. Define $\hat{P}_n = (\sum_{k=0}^{n} \hat{\phi}_k\hat{\phi}_k^T)^{-1}$, then

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \hat{P}_n\hat{\phi}_n(y_{n+1} - \hat{\phi}^T_n\,\hat{\theta}_n)$$

where

$$\hat{P}_n = \frac{\hat{P}_{n-1}}{1 + \hat{\phi}^T_n\hat{P}_{n-1}\hat{\phi}_n}.$$

The PEE: Since $E\,(w_{k+1} \mid y_s, \phi_s, s \leq k) = 0$, Assumption 6 holds. Hence, as shown in 2.2.5.a, the PEE is just the LSE. However, since $\phi_k = (u_k, \varepsilon_k)$ is not available due to ε_k, we need to use the recursive prediction error method (RPEM) as follows (See 2.2.5.):

Let $\zeta_k := \varepsilon_k$ and let $\phi_k := (u_k, \zeta_k(\hat{\theta}_{k-1}))$. We recursively obtain φ_k and $\hat{\zeta}_k$ as below.

$$\varphi^T_{n+1} = \hat{c}(n)\,\varphi^T_n + \phi^T_n,$$

and

$$\hat{\zeta}_{n+1}(\hat{\theta}_n) = y_{n+1} - \phi^T_n\,\hat{\theta}_n,$$

where $\hat{\theta}_n = (-\hat{a}(n),\ -\hat{c}(n))^T$. Then,

$$\hat{\theta}_{n+1}(\hat{\theta}_n) = \hat{\theta}_n + \hat{P}_{n+1}\varphi_{n+1}\hat{\zeta}_{n+1}(\hat{\theta}_n)$$

where $\hat{P}_{n+1}$ is obtained recursively by

$$\hat{P}_{n+1} = \frac{\hat{P}_n}{1 + \varphi^T_{n+1}\,\hat{P}_n\,\varphi_{n+1}}\ .$$

Simulation: (See **Figure 1b**.)

The simulation has been done using the ELS and the RPEM for parameters a and c, with the true parameter values $a^o = 0$ and $c^o = 0.8$. The system is the same as in Example 1(a); i.e.,

$$y_{k+1} = -(0)u_k - 0.8\varepsilon_k + \varepsilon_{k+1},$$

where $\{\varepsilon_k\} \sim N(0,1)$, $\{u_k\} \sim N(1,1)$. However, here we define $\theta^o := (-a^o, \ -c^o)$ so that we have

$$y_{k+1} = \theta^o \ u_k + w_{k+1},$$

where $w_{k+1} = \varepsilon_{k+1}$. The combinations of two different initial parameter values, $\hat{\theta}_0 = (-0.5, -0.3)$ and $(0, -0.8)$; and two different initial values for $P_0 = \rho I$, $\rho = 0.01$ and 1.0, were used.

For both estimators, small ρ-value seems to give a steady convergence. However, with a large ρ-value, the ELS and the RPEM usually converge with different speeds, and the estimator which converges quicker depends on the data generated during run-time on each implementation.

Example 2: Suppose that

$$y_k + a \ y_{k-1} = b \ u_{k-1} + \varepsilon_k, \quad k = 0, 1, \ldots$$

where y_k, $u_k \in R$; $\{u_k\}$ and $\{\varepsilon_k\}$ are independently normally distributed with the mean 0 and the variance 1; and $\{\varepsilon_k\}$ is independent of $\{u_k\}$.

Let $\phi_k := (y_k, \ u_k)^T$, $\theta := (-a, \ b)^T$, and $w_k := \varepsilon_k$. Then the system can be expressed as

$$y_{k+1} = \phi_k^T \theta + w_{k+1}.$$

The MLE: Note that Assumption 4 holds since $\{w_k\}$ is independently normally distributed with the mean 0 and the variance 1. Also, since y_k is independent of $\{w_s, \ s \geq k+1\}$ for $k = 0, 1, \ldots$, so is ϕ_k; i.e., Assumption 2 holds. Now,

$$\begin{aligned} p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k) &= p^{\theta}_{k|k-1}(u_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k) \\ &= p^{\theta}_{k|k-1}(\varepsilon_{k+1} \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k) \end{aligned}$$

since $u_k = (y_{k+1} + a \ y_k - \varepsilon_{k+1})/b$. Hence, $p^{\theta}_{k|k-1}(\phi_k \mid \phi_0, \ldots, \phi_{k-1}, y_1, \ldots, y_k)$ does not depend on θ; i.e., Assumption 3 holds. Therefore, as shown in 2.2.1.b, the MLE coincides with the LSE.

The LSE: Using Equation (4) in 2.2.2., we get

$$\hat{\theta}_n = (\sum_{k=0}^{n-1} \phi_k \phi_k^T)^{-1} \sum_{k=0}^{n-1} \phi_k y_{k+1}$$

Also, the recursive estimator is

$$\hat{\theta}_{n+1} = \hat{\theta}_n + P_n \phi_n (y_{n+1} - \phi_n^T \hat{\theta}_n)$$

where

$$P_n = \frac{P_{n-1}}{1 + \phi_n^T P_{n-1} \phi_n}$$

as shown in Equations (8) and (9).

The PEE: Since

$$E\,(w_{k+1} \mid y_s,\, \phi_s,\, s \leq k) = E\,(\varepsilon_{k+1}) = 0,$$

Assumption 5 holds. Hence, according to 2.2.5.a, the PEE coincides with the LSE.

Simulation: (See **Figures 2a** and **2b**)

The simulation has been done using the LSE for parameter (a, b) with the true parameter value $\theta^o = (a^o, b^o) = (0.8, 1.0)$. The system is

$$y_k - 0.8y_{k-1} = 1.0u_{k-1} + \varepsilon_k, \;\; k = 0,\, 1,\, \ldots$$

where $\{u_k\} \sim N(0,1)$ and $\{\varepsilon_k\} \sim N(0,1)$. Let $P_0 = \rho I$. The behavior of the estimators change according to the values for P_0.

For $\hat{\theta}_0 = (0, 0)$, given $\rho \geq 1$, the parameter estimates converge quickly especially at the early stage. (See **Figure 2a**.) However, when the true value of the parameter is given as the initial value, the behavior of the parameter estimates that were obtained with large ρ-value kept changing in almost the same manner as when $\hat{\theta}_0$ was different than the true value. On the other hand, the parameter estimates obtained with small ρ-values stayed close to the true parameter value when the true parameter value is given [2].

References

[1] P.R. Kumar and P. Varaiya (1986), Stochastic Systems: Estimation, Identification, and Adaptive Control.

[2] L. Ljung and T. Söderström (1983), Theory and Practice of Recursive Identification

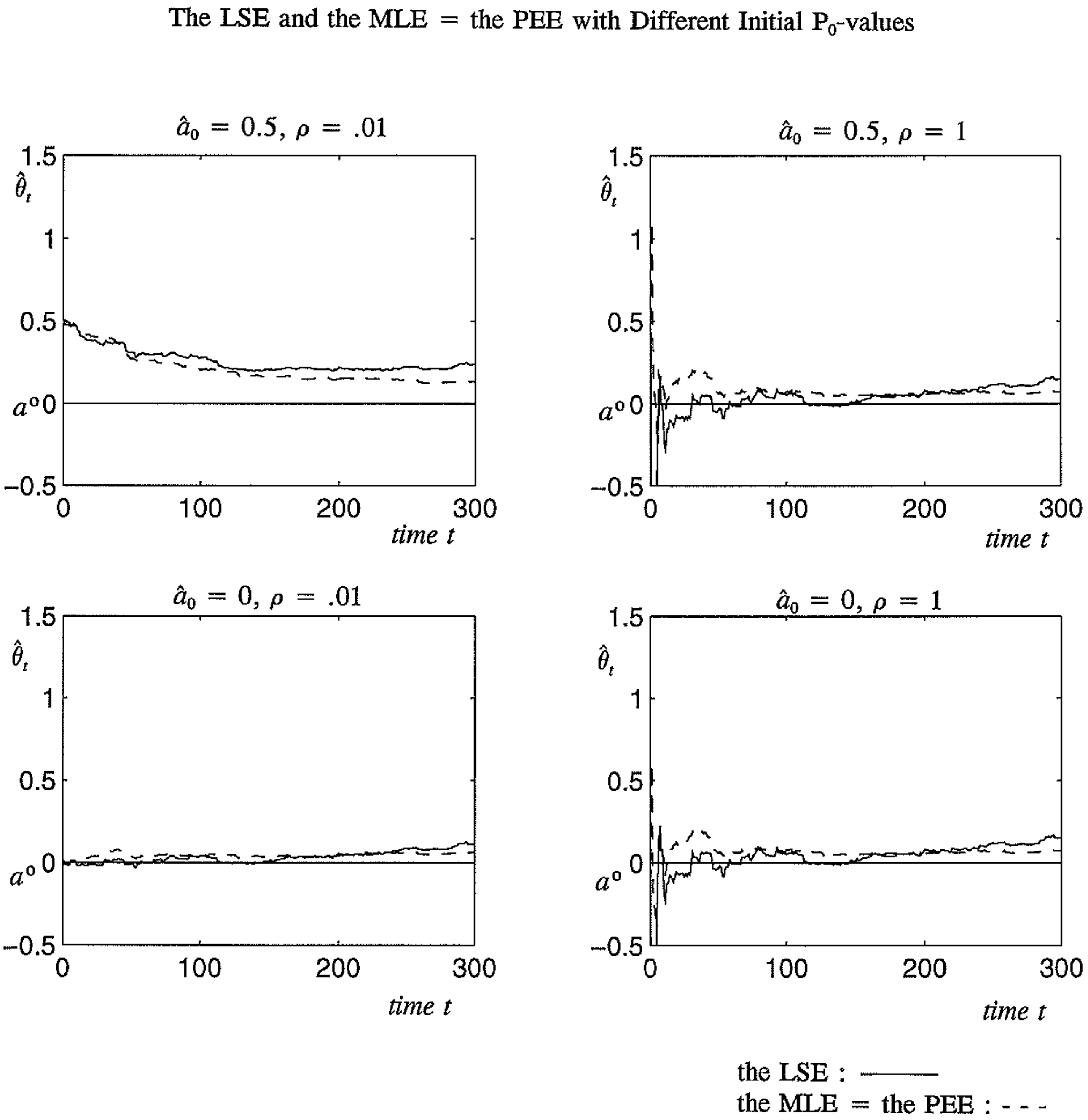

Figure 1.a

Parameter estimates for Example 1(a). $P_0 = \rho$.

The ELS and the RPEM with Different Initial P_0-values

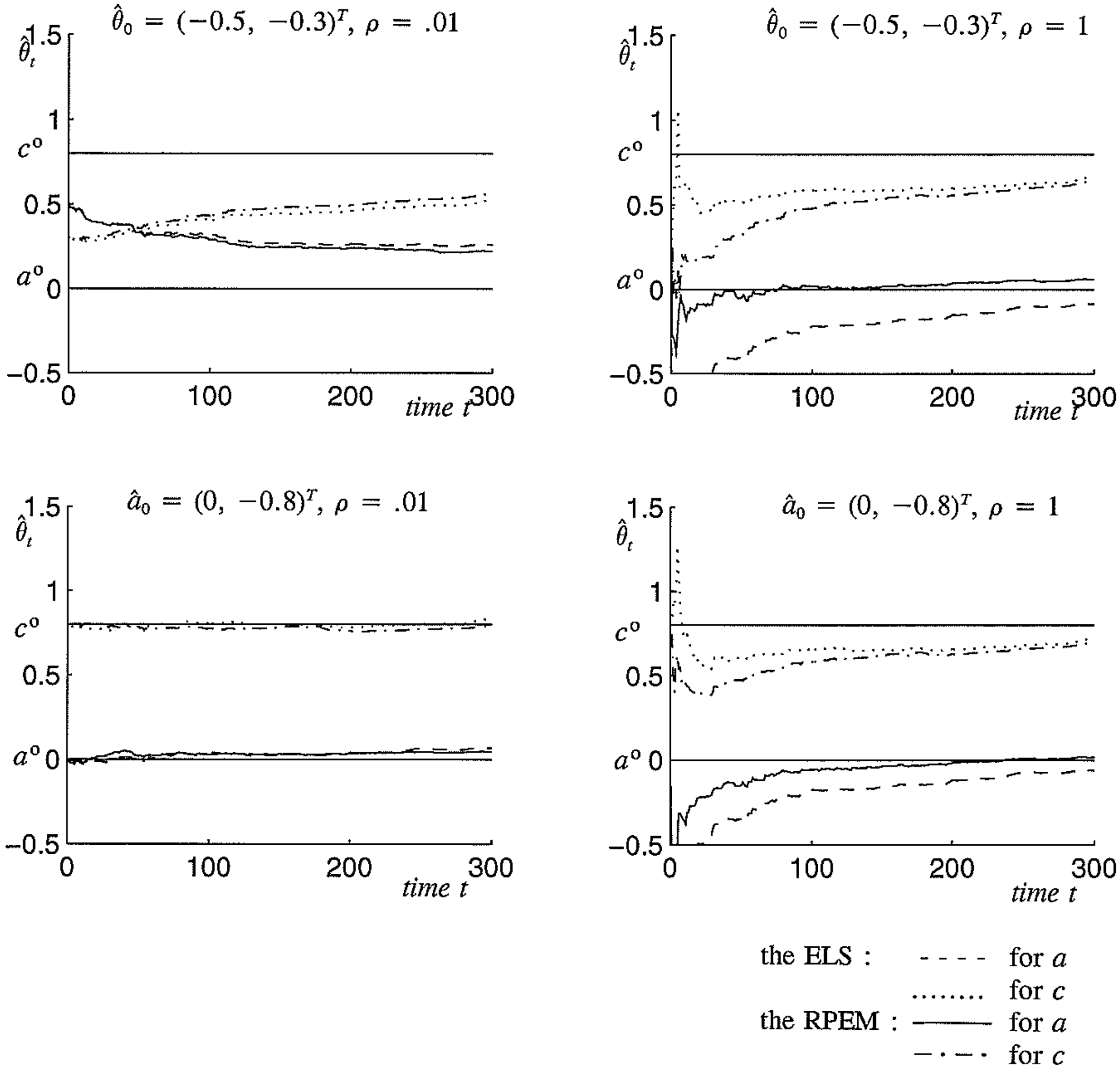

Figure 1.b

Parameter estimates for Example 1(b). $P_0 = \rho I$, $\theta^o = (-a^o,\ -c^o)^T = (0,\ -0.8)^T$.

The LSE with Different Initial P_0-values

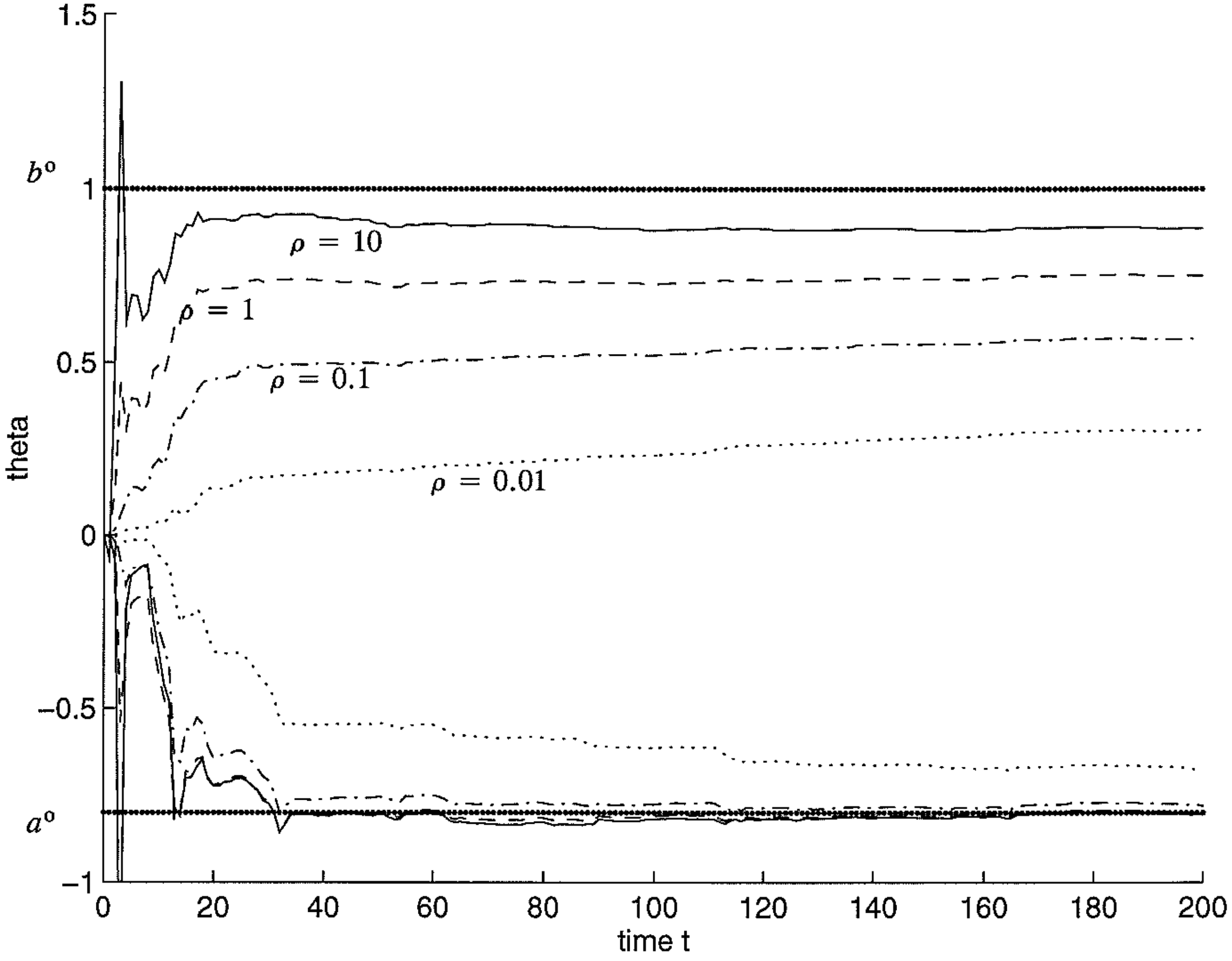

Figure 2.a

Parameter estimates for Example 2. Initial values were $\hat{\theta}_0 = (0, 0)^T$, $P_0 = \rho I$.

The LSE with Different Initial P_0-values

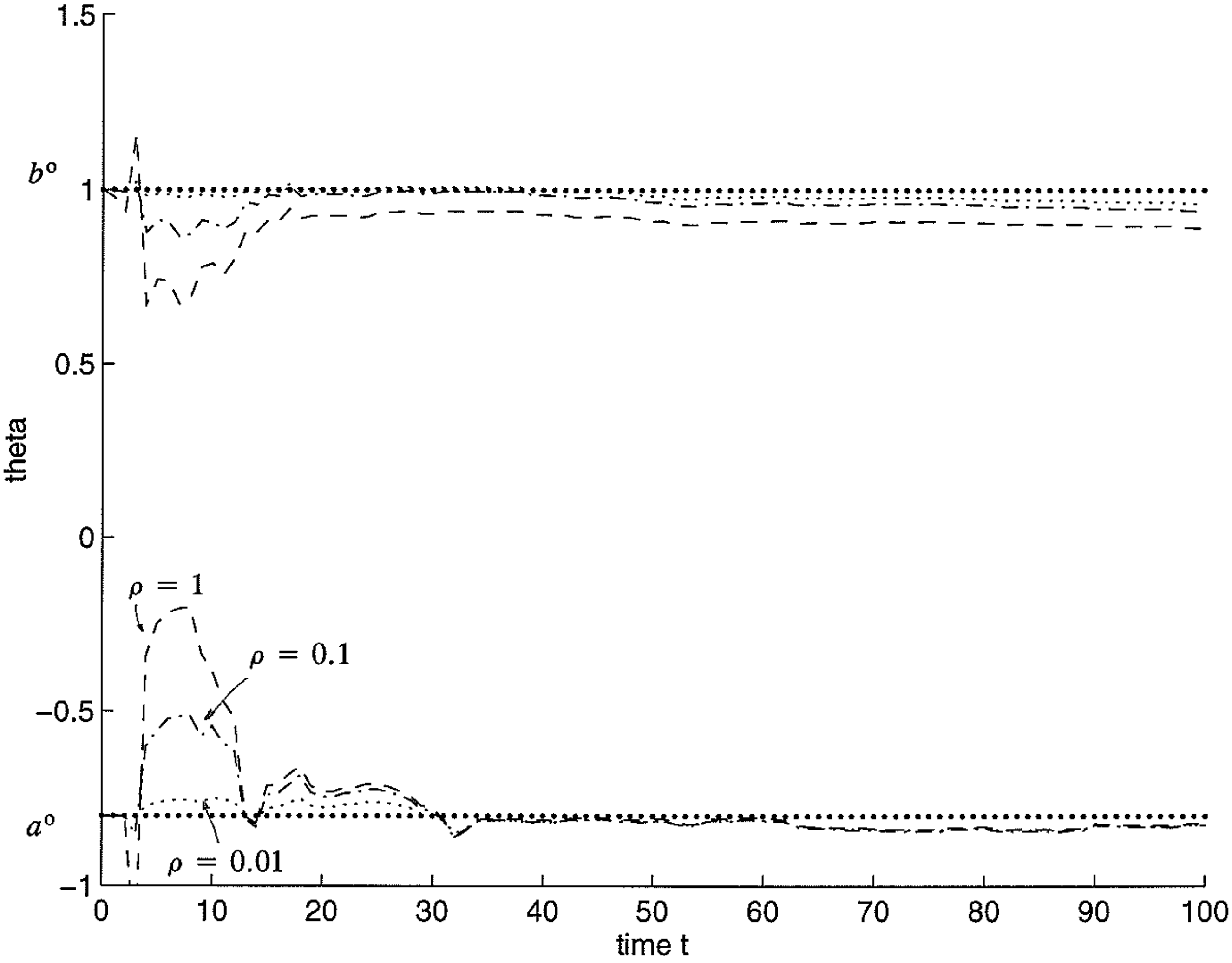

Figure 2.b

Parameter estimates for Example 2. Initial values were $\hat{\theta}_0 = (0.8,\ 1.0)^T$, $P_0 = \rho I$.

Attachments

```
                       %  ***** Example 1 ***** %
%
%  This program simulates the system
%
%  y(k+1) = theta*u(k) + w(k+1), k = 0, 1, ...
%
%  where
%
%  w(1) = v(1),
%  w(k+1) = -c*v(k) + v(k+1), k = 1, 2, ...,
%  y(k), u(k) are scalars,
%  {u(k)} is N(muU,cigmaU) and is iid,
%  {v(k)} is N(0,cigmaV) and is iid, and
%  {v(k)} is independent of {u(k)}.

% ------------------------------------------------------------------ %
%  Initialization

t0 = input('** Input initial time  =>   ');
tf = input('** Input final time  =>   ');
dt = input('** Input time step dt  =>   ');
t = t0:dt:tf;
npts = length(t) - 1;

y = zeros(1,npts);

muU = input('** Input the mean of u =>  ');
varU = input('** Input the variance of u =>  ');
u = randn(1,npts)*sqrt(varU) + muU;

varV = input('** Input the variance of v =>  ');
v = randn(1,npts)*sqrt(varV);

a = input('** Input a constant a =>  ');
c = input('** Input a constant c =>  ');

% ------------------------------------------------------------------ %
%  Simulation: of the real system.
% ------------------------------------------------------------------ %
%
%  Note: u(k) and theta#(:,k) represent u(k-1) and theta(k-1),
%        respectively, in the formula in paper.

y(1) = -a*u(1) + v(1);
for k=1:npts-1,
   y(k+1) = -a*u(k+1) - c*v(k) + v(k+1);
end

% ------------------------------------------------------------------

fprintf('\n-- For Estimation --\n');

% ------------------------------------------------------------------ %
% ------------------------------------------------------------------ %
%  (a)  c is known:

      theta = -a;

clg;
tt = t(1:npts+1);
```

```
for k=1:2

   if k == 1
      a0 = input('** Input a(0) for the 1st case =>  ');
   else
      a0 = a;
   end

   for j=1:2

      P0 = input('** Input P(0) for the 1st case =>  ');

      % ------------------------------------------------------------- %
      %  Recursive estimation of theta, using the LSE
      % ------------------------------------------------------------- %

      theta1 = zeros(1,npts+1);
      theta1(1) = -a0;
      theta1(2) = theta1(1) + P0*u(1)*(y(1)-u(1)*theta1(1));

      P = P0;
      for n = 1:npts-1
         P = P/(1 + P * u(n+1)^2);
         theta1(n+2) = theta1(n+1) + P*u(n+1) * (y(n+1) - theta1(n+1) * u(n+1));
      end

      % ------------------------------------------------------------- %
      %  Recursive estimation of theta, using the MLE = the PEE
      % ------------------------------------------------------------- %

      theta2 = zeros(1,npts+1);

      P = P0;
      uu = u(1);
      yy = y(1);
      theta2(1) = -a0;
      theta2(2) = theta2(1) + P*uu*(yy-uu);

      for n = 1:npts-1
         uu = c*uu + u(n+1);
         yy = c*yy + y(n+1);
         P = P/(1 + P * uu^2);
         theta2(n+2) = theta2(n+1) + P*uu*(yy-theta2(n+1)*uu);
      end

      if k+j == 2
         subplot(2,2,1);
      elseif k == 1 & j == 2
         subplot(2,2,2);
      elseif k == 2 & j == 1
         subplot(2,2,3);
      else
         subplot(2,2,4);
      end
      plot(tt,a-tt+tt,'-',tt,-theta1,'-',tt,-theta2,'--');
      axis([0 npts -.5 1.5]);
   end
end

print ex1temp.ps;
```

```
% ------------------------------------------------------------------- %
% ------------------------------------------------------------------- %
%  (b)  c is unknown:

clg;
for k=1:2

   if k == 1
      a0 = input('** Input a(0) for the 2nd case =>  ');
      c0 = input('** Input c(0) for the 2nd case =>  ');
   else
      a0 = a;
      c0 = c;
   end
   theta00 = [-a0, -c0]';

   for j=1:2

      P0 = input('** Input P(0) for the 2nd case =>  ');

      % ------------------------------------------------------------- %
      %  Recursive estimation of theta, using the RPEM
      % ------------------------------------------------------------- %

      theta3 = zeros(2,npts+1);

      P = P0*eye(2);
      theta3(:,1) = theta00;
      phi = [u(1),0]';
      dd = phi;
      d = y(1) - phi'*theta3(:,1);
      P = P/(1+dd'*P*dd);
      theta3(:,2) = theta3(:,1) + P*dd*d;
      d = y(1) - phi'*theta3(:,2);

      for n = 1:npts-1
         phi = [u(n+1),d]';
         dd = (-theta3(2,n+1)*dd' + phi')';
         d = y(n+1) - phi'*theta3(:,n+1);
         P = P/(1+dd'*P*dd);
         theta3(:,n+2) = theta3(:,n+1) + P*dd*d;

         d = y(n+1) - phi'*theta3(:,n+2);
      end

      % ------------------------------------------------------------- %
      %  Recursive estimation of theta, using the ELS
      % ------------------------------------------------------------- %

      theta4 = zeros(2,npts+1);

      theta4(:,1) = theta00;
      phi = [u(1),0]';
      P = P0*eye(2);
      theta4(:,2) = theta4(:,1) + P*phi*(y(1) - phi'*theta4(:,1));

      for n = 1:npts-1
         ee = y(n) - phi'*theta4(:,n+1);
         phi = [u(n+1),ee]';
         P = P/(1 + phi'*P*phi);
         theta4(:,n+2) = theta4(:,n+1) + P*phi*(y(n+1) - phi'*theta4(:,n+1));
      end
      if k+j == 2
         subplot(2,2,1);
      elseif k == 1 & j == 2
         subplot(2,2,2);
      elseif k == 2 & j == 1
         subplot(2,2,3);
      else
         subplot(2,2,4);
      end
      axis([0 npts -.5 c*1.5]);

      hold;
      plot(tt,a-tt+tt,’-’,tt,c-tt+tt,’-’);
      plot(tt,-theta3(1,:),’-’,tt,-theta3(2,:),’-.’);
      plot(tt,-theta4(1,:),’--’,tt,-theta4(2,:),’:’);
      hold off;

   end
end
```

```
                       %  ***** Example 2 ***** %
%
%       This program simulates the system
%
%  y(k) + a*y(k-1) = b*u(k-1) + e(k), k = 0, 1, ...
%
%  where
%
%  y(k), u(k) are scalars,
%  {u(k)} is N(0,1) and is iid,
%  {e(k)} is N(0,1) and is iid, and
%  {u(k)} is independent of {e(k)}.
%       Let
%           phi(k) = (y(k),u(k))',
%           theta = (-a,b)', and
%           w(k) = e(k)
%  so that we have a system
%           y(k) = phi(k-1)'*theta + w(k)
%
% ------------------------------------------------------------------- %
%  Initialization

t0 = input('** Input initial time  =>   ');
tf = input('** Input final time  =>   ');
dt = input('** Input time step dt  =>   ');
t = t0:dt:tf;
npts = length(t) - 1;

y = zeros(1,npts);
u = randn(1,npts);
w = randn(1,npts);

trueTheta = input('** Input the true parameter vector theta =>  ');
trueTheta = trueTheta';

fprintf('\n-- For Estimation --\n');

y0 = input('** Input y(0) =>  ');

% ------------------------------------------------------------------- %
%  Simulation: of the real system.
% ------------------------------------------------------------------- %
%  Note: u(k) represents u(k-1) in the formula in paper.

phi = [y0, u(1)]';
y(1) = phi'*trueTheta + w(1);
for k=1:npts-1
   phi = [y(k),u(k+1)]';
   y(k+1) = phi'*trueTheta + w(k+1);
end

tt = t(1:npts+1);
```

```
clg;
axis([0 npts -1 1.5]);
hold;
plot(tt,-trueTheta(1)-tt+tt,’-’,tt,trueTheta(2)-tt+tt,’-’);
done = 0;
i = 1;
while ~done
   alpha = input(’** Input alpha for P(0) =>  ’);
   P0 = alpha*eye(2);

   % --------------------------------------------------------------- %
   %  Recursive estimation of theta, using the LSE with theta0 = [0,0]
   % --------------------------------------------------------------- %

   %  Note: theta(:,k) represents theta(k-1) in the formula in paper.

   theta = zeros(2,npts+1);

   theta(:,1)  = [0,0]’;
   phi = [y0, u(1)]’;
   theta(:,2)  = theta(:,1)  + P0*phi*(y(1)-phi’*theta(:,1));

   P = P0;
   for n = 1:npts-1
      phi = [y(n),u(n+1)]’;
      P = P/(1 + phi’*P*phi);
      theta(:,n+2)  = theta(:,n+1)  + P*phi*(y(n+1)-phi’*theta(:,n+1));
   end

   if i == 1
      str = ’:’;
   elseif i == 2
      str = ’-.’;
   elseif i == 3
      str = ’--’;
   else i >= 4
      str = ’-’;
   end
   plot(tt,-theta(1,:),str,tt,theta(2,:),str);
   i = i+1;

   done = input(’** Input 1 if done, 0 if not =>  ’);
end
xlabel(’time t’); ylabel(’theta’);

hold off;
print ex2temp.ps;
```

```
clg;
axis([0 npts -1 1.5]);
hold;
plot(tt,-trueTheta(1)-tt+tt,'-',tt,trueTheta(2)-tt+tt,'-');
done = 0;
i = 1;
while ~done
   alpha = input('** Input alpha for P(0) =>  ');
   P0 = alpha*eye(2);

   % --------------------------------------------------------------- %
   %  Recursive estimation of theta, using the LSE with theta0 = the
   %  true theta value.
   % --------------------------------------------------------------- %

   %  Note: theta(:,k) represents theta(k-1) in the formula in paper.

   theta = zeros(2,npts+1);

   theta(:,1) = trueTheta;
   phi = [y0, u(1)]';
   theta(:,2) = theta(:,1) + P0*phi*(y(1)-phi'*theta(:,1));

   P = P0;
   for n = 1:npts-1
      phi = [y(n),u(n+1)]';
      P = P/(1 + phi'*P*phi);
      theta(:,n+2) = theta(:,n+1) + P*phi*(y(n+1)-phi'*theta(:,n+1));
   end

   if i == 1
      str = ':';
   elseif i == 2
      str = '-.';
   elseif i == 3
      str = '--';
   else i >= 4
      str = '-';
   end
   plot(tt,-theta(1,:),str,tt,theta(2,:),str);
   i = i+1;

   done = input('** Input 1 if done, 0 if not =>  ');
end

xlabel('time t'); ylabel('theta');
hold off;
```