The Annals of Mathematical Statistics
1970, Vol. 41, No. 2, 722–726

A DUALITY BETWEEN AUTOREGRESSIVE AND MOVING AVERAGE PROCESSES CONCERNING THEIR LEAST SQUARES PARAMETER ESTIMATES[1]

BY DAVID A. PIERCE

University of Missouri

1. Introduction. The methods employed in least squares parameter estimation for moving average (MA) processes differ from those appropriate for autoregressive (AR) processes, as only the latter are linear in the parameters. There is nevertheless an interesting duality between these two classes of time series models: if AR and MA series, each of the same order and with the same parameter values, are generated from the same sequence of errors, then to a close approximation the least squares estimates calculated from the MA series will *under*estimate the true parameter values by the same amount that those determined from the AR series will *over*-estimate them. This relation is established in Section 3 via a linear approximation of the moving average errors (considered as functions of the parameters for a given series) in a neighborhood of the true parameter values. In Section 4 the large-sample distribution of the estimates in any MA process is then obtained as a direct consequence of this relation and of known results for AR processes, and some properties of the least squares estimates in both classes of processes are examined.

Let us introduce notation and terminology to be employed. A sequence $\{x_t\}$ follows an AR *process of order p* if it is governed by the relation

$$x_t = \sum_{j=1}^{p} \theta_j x_{t-j} + a_t \tag{1.1}$$

where the $\{a_t\}$ are $N(0, \sigma^2)$ and independent, and the parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ are such that the roots of the auxiliary equation

$$\theta(u) = 1 - \theta_1 u - \cdots - \theta_p u^p = 0 \tag{1.2}$$

lie outside the unit circle (the set of all $\boldsymbol{\theta}$ satisfying this condition will be referred to as the *admissible parameter space*). Following [3] we may define a *backward shift operator B* by the relation $Bw_t = w_{t-1}$ for any sequence $\{w_t\}$, and equation (1.1) may then be written

$$\theta(B)x_t = a_t \tag{1.3}$$

where $\theta(B) = (1 - \sum_{j=1}^{p} \theta_j B^j)$ analogous to (1.2).

In the MA process the recursive relationship is in terms of the deviates $\{a_t\}$ rather than the observations themselves; a sequence $\{y_t\}$ is a moving average of order p if

$$y_t = -\sum_{j=1}^{p} \theta_j a_{t-j} + a_t = \theta(B)a_t \tag{1.4}$$

with $\{a_t\}$ and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ as before.

Received April 28 1969.

[1] Research begun at the University of Wisconsin and supported in part by the Wisconsin Alumni Research Foundation.

2. Least squares estimation. The equations (1.3) and (1.4) are well defined for any $t = \cdots, -1, 0, 1, \cdots$; however, in practice we will have available a sample series [say $x_1, x_2, \cdots, x_n$ or $y_1, y_2, \cdots, y_n$], and our problem is then to estimate the unknown parameters $\boldsymbol{\theta}$. The method of least squares leads us to the minimization of the sum of squares

$$S(\mathring{\boldsymbol{\theta}}) = \textstyle\sum_t [a_t(\mathring{\boldsymbol{\theta}})]^2 \tag{2.1}$$

as a function of the indeterminate $\mathring{\boldsymbol{\theta}}$. In (2.1), the quantity $a_t(\mathring{\boldsymbol{\theta}})$ is obtained by solving either (1.3) or (1.4) for $a_t = a_t(\boldsymbol{\theta})$ which then becomes (given $\{x_t\}$ or $\{y_t\}$) a function of the point $\boldsymbol{\theta} = \mathring{\boldsymbol{\theta}}$. There is of course a need to determine or have available a set of "starting values" [$x_0, x_{-1}, \cdots, x_{-p+1}$ for (1.3) or $a_0, a_{-1}, \cdots, a_{-p+1}$ for (1.4)]; however, the effect of any particular choice of these values becomes negligible asymptotically, and we can thus ignore considerations pertaining to the initialization of the process in the following discussion.

The actual determination of the quantity $\hat{\boldsymbol{\theta}}$ which minimizes the sum of squares (2.1) is facilitated in AR processes by the fact that Equation (1.3) is linear in the parameters [$\partial a_t(\mathring{\boldsymbol{\theta}})/\partial \mathring{\theta}_j = -x_{t-j}$ independently of $\mathring{\boldsymbol{\theta}}$], whereas (1.4) is not. Thus (1.3) can be written in matrix form as

$$\mathbf{x} = X\boldsymbol{\theta} + \mathbf{a} \tag{2.2}$$

where $\mathbf{a} = (a_1, \cdots, a_n)'$, $\mathbf{x} = (x_1, \cdots, x_n)'$, and X is an $(n \times p)$ matrix whose (tj)th element is the lagged observation x_{t-j}. The least squares estimates $\hat{\boldsymbol{\theta}}^{(1)}$ of $\boldsymbol{\theta}$ for the AR series are then

$$\hat{\boldsymbol{\theta}}^{(1)} = (X'X)^{-1}X'\mathbf{x} \tag{2.3}$$

with a large-sample distribution very similar to that of the estimates in the fixed-x linear normal regression model [7, Section 4].

The situation is fundamentally different for MA processes, however; letting

$$\pi(u) = 1 - \textstyle\sum_{j=1}^{\infty} \pi_j u^j \tag{2.4}$$

be the solution of $\pi(u)\theta(u) = 1$ [whose existence is guaranteed by the admissibility of $(\theta_1, \cdots, \theta_p)$], equation (1.3) becomes

$$[\theta(B)]^{-1} y_t = \pi(B) y_t = a_t,$$

and thus $a_t(\mathring{\boldsymbol{\theta}}) = [\mathring{\theta}(B)]^{-1} y_t$, whose derivative with respect to $\mathring{\theta}_j$ is a function of $\mathring{\boldsymbol{\theta}}$. Thus nonlinear estimation methods are employed in practice to obtain the least squares estimates, say $\hat{\boldsymbol{\theta}}^{(2)}$, of $\boldsymbol{\theta}$ in the MA process (1.4).

Despite this basic difference concerning linearity in these two classes of processes, however, there exists an important similarity in the behavior (asymptotically) of their least squares estimates, as summarized in Section 1. This duality is obtained in the following section by considering AR and MA processes each involving the same parameters $\boldsymbol{\theta}$ and generated from the same deviates $\{a_t\}$.

3. The correspondence between $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$. For a given sequence $\{a_t\}$ of independent $N(0, \sigma^2)$ deviates and a given vector of admissible parameters

$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$, suppose we have AR and MA series, $\{x_t\}$ and $\{y_t\}$, generated by the relations (1.3) and (1.4) respectively. In particular, then,

$$y_t = [\theta(B)]^2 x_t. \tag{3.1}$$

Suppose $\dot{\boldsymbol{\theta}}$ is a point in the admissible parameter space which is "close" (in a sense to become evident later) to the true parameter values $\boldsymbol{\theta}$. Then in terms of $\dot{\boldsymbol{\theta}}$ we can define, corresponding to (2.1),

(i) an AR error

$$\dot{a}_t^{(1)} = a_t^{(1)}(\dot{\boldsymbol{\theta}}) = \dot{\theta}(B)x_t, \tag{3.2}$$

and

(ii) an MA error

$$\dot{a}_t^{(2)} = a_t^{(2)}(\dot{\boldsymbol{\theta}}) = [\dot{\theta}(B)]^{-1}y_t. \tag{3.3}$$

Now since $\dot{a}_t^{(1)}$ is linear in $\dot{\boldsymbol{\theta}}$, we have for all $\dot{\boldsymbol{\theta}}$ the exact expression [noting that at $\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}$, $\dot{a}_t^{(1)} = \dot{a}_t^{(2)} = a_t$]

$$\begin{aligned} \dot{a}_t^{(1)} &= \sum_{j=1}^{p} (\dot{\theta}_j - \theta_j)\frac{\partial \dot{a}_t^{(1)}}{\partial \dot{\theta}_j} + a_t \\ &= \sum_{j=1}^{p} (\theta_j - \dot{\theta}_j)x_{t-j} + a_t \end{aligned} \tag{3.4}$$

or in matrix form

$$\dot{\mathbf{a}}^{(1)} = X(\boldsymbol{\theta} - \dot{\boldsymbol{\theta}}) + \mathbf{a}, \tag{3.5}$$

analogous to (2.2). The least squares estimation may then be performed directly on the "observations" $\dot{a}_1^{(1)}, \cdots, \dot{a}_n^{(1)}$ in (3.5) to give

$$\hat{\boldsymbol{\theta}}^{(1)} - \dot{\boldsymbol{\theta}} = (X'X)^{-1}X'\dot{\mathbf{a}}^{(1)}. \tag{3.6}$$

Recalling that the MA error $\dot{a}_t^{(2)}$ is not linear in $\dot{\boldsymbol{\theta}}$, the analogue of (3.4) for moving average series is an approximation via a first order Taylor expansion about $\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}$:

$$\dot{a}_t^{(2)} = \sum_{j=1}^{p} (\dot{\theta}_j - \theta_j)\left.\frac{\partial \dot{a}_t^{(2)}}{\partial \dot{\theta}_j}\right|_{\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}} + a_t \tag{3.7}$$

where the error thereby introduced is $O(|\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2)$, with $|\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}| = [\sum(\dot{\theta}_j - \theta_j)^2]^{\frac{1}{2}}$ denoting the Euclidean distance in p-space between $\dot{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. The derivatives in (3.7) are

$$\begin{aligned} \left.\frac{\partial \dot{a}_t^{(2)}}{\partial \dot{\theta}_j}\right|_{\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}} &= \left.\frac{\partial}{\partial \dot{\theta}_j}[\dot{\theta}(B)]^{-1}y_t\right|_{\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}} \\ &= \dot{\theta}^{-2}(B)B^j y_t\,|_{\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}} \\ &= \theta^{-2}(B)y_{t-j} \\ &= x_{t-j}, \end{aligned} \tag{3.8}$$

the last equality following from (3.1). Thus (3.7) becomes $\dot{a}_t^{(2)} = \sum(\dot{\theta}_j - \theta_j)x_{t-j} + a_t$ or

$$\text{(3.9)} \qquad \dot{\mathbf{a}}^{(2)} = X(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathbf{a}$$

where X is the *same* as in (3.5) and (2.2). Analogous to (3.6) we now have approximately

$$\text{(3.10)} \qquad \dot{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{(2)} = (X'X)^{-1}X'\dot{\mathbf{a}}^{(2)}.$$

The equations (3.6) and (3.10) both involve the least squares estimates $\hat{\boldsymbol{\theta}}^{(i)}$, $i = 1, 2$ and the fixed but arbitrary values $\dot{\boldsymbol{\theta}}$. If we now set $\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}$, these equations become

$$\text{(3.11)} \qquad \hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta} = (X'X)^{-1}X'\mathbf{a} \qquad \text{and}$$

$$\text{(3.12)} \qquad \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(2)} = (X'X)^{-1}X'\mathbf{a}, \qquad \text{whence}$$

$$\text{(3.13)} \qquad (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(2)}) = -(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(1)})$$

where the error of approximation in (3.13) is of order n^{-1} in probability and becomes asymptotically negligible.

Equation (3.13) shows that for large samples $\hat{\boldsymbol{\theta}}^{(2)}$ from the MA series *under*-estimates the true value $\boldsymbol{\theta}$ by the same amount that $\hat{\boldsymbol{\theta}}^{(1)}$ from the AR series *over*-estimates $\boldsymbol{\theta}$, this being for the particular instance when both series are generated from the same white noise. It follows from this that the asymptotic distribution of $\hat{\boldsymbol{\theta}}^{(2)}$ for any MA process (1.4) is the same as that of the dual AR process (1.3), since the latter distribution is known to be normal and hence symmetric. (The large-sample distribution of $\hat{\boldsymbol{\theta}}^{(2)}$ has been established via other methods by Whittle [8] and Box and Jenkins [3]; Box and Jenkins also obtain the joint distribution of the estimates in a mixed AR–MA process.) We briefly discuss this common distribution and the properties of $\hat{\boldsymbol{\theta}}^{(i)}$ in the following section.

4. Distribution and properties of the least squares estimates; effect of nonnormality of $\{a_t\}$. The large-sample distribution of the least squares estimates in an AR process was obtained by Mann and Wald [7] who showed (i) that asymptotically,

$$\text{(4.1)} \qquad n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}) \sim N(\mathbf{0}, \sigma^2\Gamma^{-1})$$

where $\Gamma = (\gamma_{|i-j|})$ is a $(p \times p)$ matrix of the *autocovariances* $\gamma_k = E(x_t x_{t+k})$ of the AR process (1.3); and (ii) that

$$\text{(4.2)} \qquad \Gamma = \operatorname{plim} n^{-1}X'X$$

with X as in (2.2). It follows from (3.13) that equation (4.1) also gives the asymptotic distribution of the estimates $\hat{\boldsymbol{\theta}}^{(2)}$ of the MA process (1.4), with the stipulation, of course, that Γ is the autocovariance matrix not of this process but of the dual AR process (1.3).

Furthermore, under the normality assumption the minimization of (2.1) is equivalent (asymptotically) to the maximization of the likelihood function, so that $\hat{\boldsymbol{\theta}}^{(i)}$ is also the ML estimate of $\boldsymbol{\theta}$, $i = 1, 2$. Thus for large samples, least squares estimates in both AR and MA processes are efficient, and the information matrices for either are given by $(1/\sigma^2)\Gamma$. In particular, the problem of efficient estimation in MA processes ([5], [6, Chapter 2], [8, page 431]) is seen to be equivalent to the problem of minimizing (2.1), which is readily accomplished with the aid of a computer and employing standard nonlinear estimation techniques [2], [4].

While the preceding discussion has been given under the assumption of a normal distribution for $\{a_t\}$, many large-sample properties of the estimates $\hat{\boldsymbol{\theta}}^{(i)}$ hold under wider conditions. Mann and Wald in their investigation obtained the distribution (4.1) for $\hat{\boldsymbol{\theta}}^{(1)}$ assuming only that the $\{a_t\}$ possessed moments of all orders, and T. W. Anderson [1] showed that the existence of moments of order two is sufficient. By (3.13) these remarks can also be made for least squares parameter estimates in MA processes. Clearly also the large sample covariance matrices of $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ will be the same (and given by (4.1)), though in general these will no longer be related to the information matrices. What is lost when the normality assumption is not satisfied is efficiency, as the ML estimates will no longer necessarily coincide with the least squares estimates. (The latter will however still have minimum asymptotic variance within a class of estimates limited in a way analogous to the restriction of linearity and unbiasedness under which the Gauss–Markov theorem holds in the non-normal regression model; Whittle [8] indicates these restrictions in terms of the spectrum.)

Let us conclude by briefly illustrating these considerations with the first order AR and MA processes,

$$x_t = \theta x_{t-1} + a_t \quad \text{and} \tag{4.3}$$

$$y_t = -\theta a_{t-1} + a_t, \tag{4.4}$$

respectively. The admissibility restriction on the parameter is that $|\theta| < 1$. If $E(a_t^2) < \infty$, then the least squares estimate of θ in either process is asymptotically normal with mean θ and variance $(1-\theta^2)/n$ [the (1×1) covariance matrix is $\Gamma = \operatorname{Var}(x_t) = \sigma^2/(1-\theta^2)$]. If in addition the $\{a_t\}$ are normally distributed then for large samples the information is $I(\theta) = \sum_0^\infty \theta^{2j} = (1-\theta^2)^{-1}$, and $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$ are efficient.

REFERENCES

[1] ANDERSON, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Ann. Math. Statist.* **30** 676–687.

[2] BOX, G. E. P. and JENKINS, G. M. (1970). *Time Series Analysis, Forecasting and Control.* Holden–Day, San Francisco. Chapter 7.

[3] BOX, G. E. P., JENKINS, G. M. and BACON, D. W. (1967). Models for forecasting seasonal and nonseasonal time series. *Spectral Analysis of Time Series*, ed. B. Harris. Wiley, New York, 271–311.

[4] DRAPER, N. R. and SMITH, H. (1966). *Applied Regression Analysis.* Wiley, New York, Chapter 10.

[5] DURBIN, J. (1959). Efficient estimation of parameters in moving average models. *Biometrika* **46** 306–316.

[6] HANNAN, E. J. (1960). *Time Series Analysis.* Methuen, London.

[7] MANN, H. B. and WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica* **11** 173–220.

[8] WHITTLE, P. (1953). Estimation and information in stationary time series. *Ark. Mat.* **2** 423–434.