NBER WORKING PAPER SERIES

CERTAIN ASPECTS OF GENERALIZED BOX-JENKINS MODELS

Richard W. Hill*

Working Paper No. 82

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

May 1975

* NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

Abstract

We define a class of models that are generalizations of regression models and moving average-autoregressive time series models. Then we investigate the asympotic and computational properties of the maximum likelihood estimator, with numerical examples. The main conclusion is that care must be excercised when using simple approximations to the covariance matrix of the estimates.

Contents

The main purpose of the paper is to examine certain aspects of Box-Jenkins models: specifically we will examine computational methods and approximations to asymptotic covariance matrices. We begin, however, by introducing a more general setup.

I. THE MODEL

We will work with the following model. Let β be a kxl vector of parameters, m a twice differentiable function m: $R^k \to R^n$, so that $m(\beta)$ is an nxl vector. $V(\theta)$ is an nxn symmetric positive definite matrix, whose elements are a function of the pxl vector θ.

Our model is

$$Y = m(\beta) + \varepsilon, \tag{1-1}$$

where $\varepsilon \sim N_n(0, \sigma^2 V(\theta))$,

so that if

$$V(\theta) = [V^{\frac{1}{2}}(\theta)]\,[V^{\frac{1}{2}}(\theta)]^T,$$

$$V^{-\frac{1}{2}}(\theta)\,\varepsilon \sim N_n(0, \sigma^2 I_n). \tag{1-2}$$

For example if $V(\theta) = I_n$, we have the usual nonlinear model, and if

$m(\beta) = X\beta$ then we have

$Y - X\beta \sim N(0, \sigma^2 I_n)$ which is the usual linear regression model.

For convenience, we put $f(\beta) = Y - m(\beta)$, so that $f(\beta)$ is the nx1 vector of residuals. We let $\gamma = \begin{pmatrix}\beta\\ \theta\end{pmatrix}$, the combined parameter vector. In our applications we will find that p, the dimension of θ, is much smaller than n, so that $V(\theta)$ is unknown only up to few parameter values, which we wish to estimate. For example, if Y were a zero mean time series, we could take $f(\beta) = Y$, and perhaps assume

$$V^{-\frac{1}{2}}(\theta) \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 \\ \theta & 1 & 0 & . & . & . & 0 \\ 0 & \theta & 1 & . & . & . & 0 \\ & . & & & & & \\ & . & & & & & \\ & . & & & & & \\ 0 & 0 & 0 & . & . & \theta & 1 \end{pmatrix} .$$

This is a one parameter model, in which we are trying to estimate the correlation between Y_i and Y_{i+1}, assuming that Y_i and Y_{i+t} are uncorrelated for $t \geq 2$. The usual Box-Jenkins models described in Box and Jenkins (1970) are special cases of (1-1). In fact, they can be written as

$$Y_i - \rho_1 Y_{i-1} - \rho_2 Y_{i-2} - \cdots - \rho_a Y_{i-a} = \varepsilon_i - \phi_1 \varepsilon_{i-1} - \cdots - \phi_b \varepsilon_{i-b}, \qquad (1\text{-}3)$$

where

$\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$ variables.

In our notation

$$P(\rho)Y = T(\phi)\,\varepsilon, \qquad (1\text{-}4)$$

where $\rho = (\rho_1, \ldots, \rho_a)^T$, $\phi = (\phi_1, \ldots, \phi_b)^T$,

$$P(\rho) = \begin{pmatrix} 1 & 0 & 0 & & & 0 \\ -\rho_1 & 1 & 0 & & & 0 \\ -\rho_2 & -\rho_1 & 1 & & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ & & & \cdots & & \\ -\rho_a & -\rho_{a-1} & -\rho_{a-2} & & & 0 \\ 0 & -\rho_a & -\rho_{a-1} & & & 0 \\ 0 & 0 & -\rho_a & & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & & & 1 \end{pmatrix} \qquad (1\text{-}5)$$

and

$$T(\phi) = \begin{pmatrix} 1 & 0 & 0 & & & 0 \\ -\phi_1 & 1 & 0 & & & 0 \\ -\phi_2 & -\phi_1 & 1 & & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ & & & \cdots & & \\ -\phi_a & -\phi_{a-1} & -\phi_{a-2} & & & 0 \\ 0 & -\phi_a & -\phi_{a-1} & & & 0 \\ 0 & 0 & -\phi_a & & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & & & 1 \end{pmatrix} \qquad (1\text{-}6)$$

Letting $$\theta = (\rho_1, \ldots, \rho_a, \phi_1, \ldots, \phi_b)^T \tag{1-7}$$

and $$V^{-\frac{1}{2}}(\theta) = T^{-1}(\phi)\, P(\rho), \tag{1-8}$$

we have $$V^{-\frac{1}{2}}(\theta)\, Y \sim N_n\,(0,\ \sigma^2\, I_n),$$

so that the Box-Jenkins models are indeed special cases of (1-1), with $m(\beta) \equiv 0$ and $V(\theta)$ given by (1-8). Throughout, we will let $P(\rho)$ and $T(\phi)$ be defined by the above matrices.

II. ESTIMATION

Since we have not assumed that $f(\beta)$ is linear, and $V(\theta)$ is not necessarily linear in θ, we are not in an exponential family, so the theory of sufficiency is not applicable. We resort to the principle of maximum likelihood. We can only observe the nxl vector Y, so we need the likelihood in terms of Y:

$$L(f,\beta,\theta,\sigma) = \frac{C \det(V^{-\frac{1}{2}}(\theta))}{\sigma^n} \exp\left[\frac{- f^T(\beta)V^{-1}(\theta)\ f(\beta)}{2\ \sigma^2}\right] , \quad (2\text{-}1)$$

where C is a constant (see Rao (1965) section 8a.4).

For all our applications we will have $\det(V(\theta)) = 1$ (see 1-5 and 1-6), so we immediately simplify things by assuming that

$$\det\left(V^{\frac{1}{2}}(\theta)\right) = 1 \qquad \text{for all values of } \theta. \quad (2\text{-}2)$$

Hence

$$\log L(f,\beta,\theta,\sigma) = - \frac{1}{2\sigma^2} f^T(\beta)\ V^{-1}(\theta)\ f(\beta) - n \log \sigma + C . \quad (2\text{-}3)$$

To maximize this we differentiate and set the derivatives to 0. (Recall that $f(\beta) = Y - m(\beta)$, so for each β, $f(\beta)$ is observable.)

$$\frac{\partial \log L}{\partial \sigma} = \frac{1}{\sigma^3} f^T(\beta)\ V^{-1}(\theta)\ f(\beta) - \frac{n}{\sigma} = 0$$

or

$$f^T(\beta)\ V^{-1}(\theta)\ f(\beta) = n\ \sigma^2 .$$

Hence

$$\hat{\sigma}^2 = \frac{f^T(\hat{\beta})\ V^{-1}(\hat{\theta})\ f(\hat{\beta})}{n} , \quad (2\text{-}4)$$

and we can treat σ^2 as a constant throughout the rest of the discussion.

Note that we are now trying to minimize

$$f^T(\beta)\ V^{-1}(\theta)\ f(\beta).$$

We write $f(\beta) = (f_1, \ldots, f_n)^T$; $V^{-1}(\theta) = (V^{ij})$ for convenience.

Then

$$\frac{\partial \log L}{\partial \beta_j} = \frac{-1}{2\sigma^2} \frac{\partial}{\partial \beta_j} \left(\sum_{ij} f_i V^{ij} f_j\right)$$

$$= \frac{-1}{2\sigma^2} \left(\sum_{ij} \left[\frac{\partial f_i}{\partial \beta_j} V^{ij} f_j + f_i V^{ij} \frac{\partial f_j}{\partial \beta_i}\right]\right)$$

$$= \frac{-1}{\sigma^2} \left(\sum_{ij} \frac{\partial f_i}{\partial \beta_j} V^{ij} f_j\right)$$

$$= \frac{-1}{\sigma^2} \left(\frac{\partial f(\beta)}{\partial \beta_j}\right)^T V^{-1}(\theta)\ f(\beta)\ .$$

$$\frac{\partial \log L}{\partial \theta_m} = \frac{-1}{2\sigma^2} \frac{\partial}{\partial \theta_m} \left(\sum_{ij} f_i V^{ij} f_j\right)$$

$$= \frac{-1}{2\sigma^2} \left(\sum_{ij} f_i \frac{\partial V^{ij}}{\partial \theta_m} f_j\right)$$

$$= \frac{-1}{2\sigma^2} f^T(\beta) \frac{\partial V^{-1}(\beta)}{\partial \theta_m} f(\beta)\ .$$

So the k+p normal equations are

$$\begin{cases} \dfrac{\partial f^T(\beta)}{\partial \beta_j} \; V^{-1}(\theta) \; f(\beta) = 0 \\ \\ f^T(\beta) \; \dfrac{\partial V^{-1}(\theta)}{\partial \theta_m} \; f(\beta) = 0 \; . \end{cases} \tag{2-5}$$

Note how the first equation imposes the usual least squares condition: residuals orthogonal (in the right metric) to the "data", represented here by the first derivative matrix.

The second equation is also an orthogonality condition, albeit somewhat less obvious: as we shall see later, for certain special cases this condition becomes more explicit.

To solve these equations we propose to use some variant of Newton's method, so we compute the second derivatives. Omitting the details we get

$$-\sigma^2 \; \frac{\partial^2 \log L}{\partial \beta_i \beta_j} = \frac{\partial f^T(\beta)}{\partial \beta_i} \; V^{-1}(\theta) \; \frac{\partial f(\beta)}{\partial \beta_j} + \frac{\partial f^T(\beta)}{\partial \beta_i \beta_j} \; V^{-1}(\theta) \; f(\beta)$$

$$-\sigma^2 \; \frac{\partial^2 \log L}{\partial \beta_i \theta_m} = \frac{\partial f^T(\beta)}{\partial \beta_i} \; \frac{\partial V^{-1}(\theta)}{\partial \theta_m} \; f(\beta) \tag{2-6}$$

$$-2\sigma^2 \; \frac{\partial^2 \log L}{\partial \theta_\ell \theta_m} = f^T(\beta) \; \frac{2V^{-1}(\theta)}{\partial \theta_\ell \theta_m} \; f(\beta) \; ,$$

and Newton's method is

$$\gamma^{(i+1)} = \gamma^{(i)} - H^{-1}(\gamma^{(i)}) \; G(\gamma^{(i)}) \; , \tag{2-7}$$

where

$$H = \begin{pmatrix} [f']^T V^{-1} f' + f'' V^{-1} f & [f']^T [V^{-1}]' f \\ [f']^T [V^{-1}]' f & f^T [V^{-1}]'' f \end{pmatrix}$$

and

$$G = \begin{pmatrix} [f']^T V^{-1} f \\ f^T [V^{-1}]' f \end{pmatrix}. \qquad (2\text{-}8)$$

The primes denoting the appropriate derivatives. (Note that the factor $\frac{1}{2}$ is omitted from the lower right hand corner of H, because it is omitted in the lower half of G). The Fisher information matrix $I(\gamma)$ is

$$E \begin{pmatrix} \frac{1}{\sigma^2} [f']^T V^{-1} f' + [f'']^T V^{-1} f & \frac{1}{\sigma^2} [f']^T [V^{-1}]' f \\ \frac{1}{\sigma^2} [f']^T [V^{-1}]' f & \frac{1}{2} \frac{1}{\sigma^2} f^T [V^{-1}]'' f \end{pmatrix}. \qquad (2\text{-}9)$$

Holland (1973) described a method for carrying out the expectation in 2-9. Since σ^2 is considered fixed, we treat it as a constant. Then

$$E \left[\frac{1}{\sigma^2} [f']^T V^{-1} f' + [f'']^T V^{-1} f \right] =$$

$$= \frac{1}{\sigma^2} [f']^T V^{-1} f' + [f'']^T V^{-1} [Ef] = \frac{1}{\sigma^2} [f']^T V^{-1} f',$$

since $f'(\beta) = m'(\beta)$ was assumed fixed;

$$E \left[\frac{1}{\sigma^2} [f']^T [V^{-1}]' f \right] = \frac{1}{\sigma^2} [f']^T [V^{-1}]' [Ef] = 0,$$

since $f(\beta) = Y - m(\beta) \sim N(0, \sigma^2 V(\theta))$, by 1-1.

$$E \left[\frac{1}{2\sigma^2} f^T [V^{-1}]'' f \right] = \frac{1}{2\sigma^2} \text{trace} \left[E [f^T [V^{-1}]'' f] \right] =$$

$$= \frac{1}{2\sigma^2} E \text{ trace } [f^T [V^{-1}]'' f] = \frac{1}{2\sigma^2} E \text{ trace } [[V^{-1}]'' f f^T] =$$

$$= \frac{1}{2\sigma^2} \text{ trace } [E[[V^{-1}]'' \, ff^T]] = \frac{1}{2\sigma^2} \text{ trace } [[V^{-1}]'' \, E[ff^T]] =$$

$$= \frac{1}{2\sigma^2} \text{ trace } [[V^{-1}]''\sigma^2 V] = \frac{1}{2} \text{ trace } [V[V^{-1}]'']$$

So we have

$$I(\gamma) = \begin{pmatrix} \frac{1}{\sigma^2}[f']^T V^{-1} f' & 0 \\ 0 & 1/2 \text{ trace } (V \frac{\partial V^{-1}}{\partial \theta_\ell \theta_m}) \end{pmatrix} . \qquad (2\text{-}10$$

Under suitable regularity conditions, it can be shown that

$$\sqrt{n}\,(\hat{\gamma}-\gamma) \xrightarrow{\mathcal{L}} N_{k+p}(0, I^{-1}(\gamma)). \qquad (2\text{-}11$$

We will assume that (2-11) holds.

We now specialize to a subset of (1-1) for which the expressions (2-7) are easy to compute.

III. THE GENERALIZED BOX-JENKINS SETUP

We restrict ourselves to the subset of (1-1) for which

$$V(\theta) = P^{-1}(\rho)\, T(\phi)\, [P^{-1}(\rho)\, T(\phi)]^T \tag{3-1}$$

so that $$V^{-\frac{1}{2}}(\theta) = T^{-1}(\phi)\, P(\rho) \tag{3-2}$$

where T, P are given by (1-5) and (1-6).

Note that det $(T(\phi))$ = det $(P(\phi))$ = 1, so that det $(V^{-\frac{1}{2}}(\theta))$ = 1 as required by (2-2).

We will use the following lemmas:

Lemma 1 Let $A(\alpha)$, $B(\alpha)$ be any non-singular matrices whose elements are a function of a scalar α. Then

i) $$\frac{\partial}{\partial\alpha} A^{-1}(\alpha) = -A^{-1}(\alpha) \frac{\partial}{\partial\alpha} A(\alpha)\, A^{-1}(\alpha)$$

ii) $$\frac{\partial}{\partial\alpha} A(\alpha)\, B(\alpha) = \frac{\partial}{\partial\alpha} A(\alpha)\, B(\alpha) + A(\alpha) \frac{\partial}{\partial\alpha} B(\alpha)$$

Proof:

i) $$I = A(\alpha)\, A^{-1}(\alpha) \ .$$

So

$$0 = \frac{\partial}{\partial\alpha} \sum_k a_{ik}(\alpha)\, a^{kj}(\alpha)$$

$$= \sum_k \left(\frac{\partial}{\partial\alpha} a_{ik}(\alpha)\, a^{kj}(\alpha) + \frac{\partial}{\partial\alpha} a^{kj}(\alpha)\, a_{ik}(\alpha)\right) \ .$$

Hence

$$0 = \frac{\partial}{\partial\alpha} A(\alpha)\, A^{-1}(\alpha) + A(\alpha) \frac{\partial}{\partial\alpha} A^{-1}(\alpha)$$

and

$$\frac{\partial}{\partial\alpha} A^{-1}(\alpha) = -A^{-1}(\alpha) \frac{\partial}{\partial\alpha} A(\alpha)\, A^{-1}(\alpha)$$

ii) follows similarly

QED

Definition A matrix A is said to be Column Triangular if A can be written as

$$A = \begin{pmatrix} a_1 & 0 & 0 & & & 0 \\ a_2 & a_1 & 0 & & & 0 \\ a_3 & a_2 & a_1 & & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ a_n & a_{n-1} & a_{n-2} & & & a_1 \end{pmatrix}$$

Lemma 2 If A and B are arbitrary column triangular matrices, then

i) AB is column triangular

ii) A^{-1} is column triangular (if it exists)

iii) AB = BA

iv) Trace $(AB^T) = \sum_{i=1}^{n} i a_{n-i}\, b_{n-i}$

v) If furthermore the entries of A are a function of the scalar α, then $\frac{\partial A}{\partial\alpha}$ is column triangular.

Proof: Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ determine A and B respectively.

i) $$(AB)_{ij} = \sum_k A_{ik} B_{kj} = \sum_{\mathcal{A}} a_{i-k+1} b_{k-j+1} ,$$

where $\mathcal{A} = \{k : 1 \leq i-k+1 \leq n \text{ and } 1 \leq k-j+1 \leq n\}$.

The range in the summation can be rewritten as

$$\begin{cases} -i \leq -k \leq n-i-1 \\ j \leq k \leq n+j-1 \end{cases}$$

or

$$\begin{cases} i \geq k \geq i+1-n \\ j \leq k \leq n+j-1 \end{cases} ,$$

so $$(AB)_{ij} = \sum_{k=j}^{i} a_{i-k+1} b_{k-j+1} .$$

Hence $$(AB)_{ij} = 0 \text{ whenever } j > i$$

and $$(AB)_{i,j+\ell} = \sum_{k=j+\ell}^{i} a_{i-k+1} b_{k-j-\ell+1}$$

$$= \sum_{j \leq k-\ell \leq i-\ell} a_{i-k+1} b_{k-j-\ell+1}$$

$$= \sum_{m=j}^{i-\ell} a_{i-m-\ell+1} b_{m-j+1}$$

$$= (AB)_{i-\ell,j} ,$$

which establishes column triangularity.

ii) Note that a column triangular matrix is lower triangular, so its inverse can be computed column by column by simple forward substitution. Letting

$$\delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases},$$

we have

$$A^{ij} = \begin{cases} \dfrac{\delta_{ij} - \sum_{k=j}^{i-1} A_{ik} A^{kj}}{A_{ii}} & \text{for } i=j, \ldots, n \\ 0 & \text{for } i=1, \ldots, j-1 \end{cases}$$

where A^{ij} is the i,j element of A^{-1}.

But $A_{ik} = a_{i-k+1}$,

so $$A^{ij} = \frac{\delta_{ij} - \sum_{k=j}^{i-1} a_{i-k+1} A^{kj}}{a_1}, \quad i=j, \ldots, n$$

$$A^{i,j+1} = \frac{\delta_{i,j+1} - \sum_{k=j+1}^{i-1} a_{i-k+1} A^{k,j+1}}{a_1}, \quad i=j+1, \ldots, n$$

hence $$A^{i+1,j+1} = \frac{\delta_{i+1,j+1} - \sum_{k=j+1}^{i} a_{i+1-k+1} A^{k,j+1}}{a_1}, \quad i=j, \ldots,$$

$$\text{and} \qquad A^{i+1,i+1} = \frac{\delta_{ij} - \sum_{k=j}^{i-1} a_{i+1-k} A^{k+1,j+1}}{a_1}, \quad i=j, \ldots, n-1$$

Since this formula allows us to successively compute the elements $A^{i+1,j+1}$ for $i=j, \ldots, n-1$ and we know that $A^{i+1,j+1} = 0$ for $i=1, \ldots, j-1$, by comparison with the formula for A^{ij} we see that $A^{i+1,j+1} = A^{ij}$ for $i=j, \ldots, n-1$, which establishes column triangularity.

iii) $$(AB)_{ij} = \sum_{k=j}^{i} a_{i-k+1} b_{k-j+1}$$

$$= \sum_{\ell=j}^{i} b_{i-\ell+1} a_{\ell-j+1} = (BA)_{ij} ,$$

by setting $\ell = i+j-k$, so that $k = i+j-\ell$.

iv) $$(AB^T)_{ii} = \sum_{k=1}^{i} A_{ik} B_{ik} = \sum_{k=1}^{i} a_{i-k+1} b_{i-k+1}$$

$$\text{Trace } (AB^T) = \sum_{i=1}^{n} (AB^T)_{ii} = \sum_{i=1}^{n} \sum_{k=1}^{i} a_{i-k+1} b_{i-k+1}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{i} a_k b_k = \sum_{k=1}^{n} \sum_{i=0}^{n-k} a_k b_k$$

$$= \sum_{k=1}^{n} (n-k+1) a_k b_k = \sum_{i=1}^{n} i a_{n-i+1} b_{n-i+1}$$

v) obvious.

QED

Note that $P(\rho)$ and $T(\phi)$ are both column triangular, so from the above lemma, we never need to actually carry around P and T, but only their first columns. This allows us to simplify things considerably. Specifically (omitting the arguments)

$$V^{-1} = [T^{-1}P]^T\, T^{-1}P$$

so
$$G = \begin{pmatrix} [T^{-1}P\ f']^T\, T^{-1}P\ f \\ [(T^{-1}P)'\ f]^T\, (T^{-1}P)\ f \end{pmatrix} \tag{3-3}$$

since $\quad [V^{-1}]' = [(T^{-1}P)']^T\, T^{-1}P + [T^{-1}P]^T\, (T^{-1}P)'$

and $\quad f^T\, (V^{-1})'\, f$ is a scalar.

Also $H =$

$$\begin{pmatrix} [T^{-1}Pf']^T\, T^{-1}Pf' + [T^{-1}Pf'']^T\, T^{-1}Pf & [(T^{-1}P)f']^T\, (T^{-1}P)'f + [(T^{-1}P)'f']^T\, (T^{-1}P)f \\ [(T^{-1}P)f']^T\, (T^{-1}P)'f + [(T^{-1}P)'f']^T\, (T^{-1}P)f & [(T^{-1}P)'f]^T\, (T^{-1}P)'f + [(T^{-1}P)''f]^T\, (T^{-1})f \end{pmatrix} \tag{3-4}$$

since $\quad (V^{-1})'' = [(T^{-1}P)'']^T\, T^{-1}P + [(T^{-1}P)']^T\, (T^{-1}P)'$.

Further simplification is possible noting that

$$\begin{cases} \dfrac{\partial}{\partial \rho_i}\ (T^{-1}P) = \dfrac{\partial\ T^{-1}}{\partial\ \rho_i}\ P + T^{-1}\, \dfrac{\partial\ P}{\partial \rho_i} = T^{-1}\, \dfrac{\partial\ P}{\partial \rho_i} = \dfrac{\partial\ P}{\partial \rho_i}\, T^{-1} \\ \dfrac{\partial}{\partial \phi_i}(T^{-1}P) = \dfrac{\partial\ T^{-1}}{\partial \phi_i}\, P = -T^{-1}\, \dfrac{\partial\ T}{\partial \phi_i}\, T^{-1}P = \dfrac{-\partial T}{\partial \phi_i}\, T^{-2}P \end{cases} \tag{3-5}$$

$$\begin{cases} \dfrac{\partial^2(T^{-1}P)}{\partial\rho_i\rho_j} = \dfrac{\partial T^{-1}}{\partial\rho_i}\dfrac{\partial P}{\partial\rho_i} + T^{-1}\dfrac{\partial^2 P}{\partial\rho_i\rho_j} = 0 \\ \dfrac{\partial^2(T^{-1}P)}{\partial\phi_i\rho_j} = \dfrac{\partial T^{-1}}{\partial\phi_i}\dfrac{\partial P}{\rho_j} = -T^{-1}\dfrac{\partial T}{\partial\phi_i}T^{-1}\dfrac{\partial P}{\partial\rho_j} = \dfrac{\partial T}{\partial\phi_i}\dfrac{\partial P}{\partial\rho_j}T^{-2} \\ \dfrac{\partial^2(T^{-1}P)}{\partial\phi_i\phi_j} = \dfrac{\partial}{\partial\phi_j}\left(-\dfrac{\partial T}{\partial\phi_i}T^{-1}T^{-1}P\right) = -\dfrac{\partial T}{\partial\phi_i}\dfrac{\partial T^{-1}}{\partial\phi_j}T^{-1}P - \dfrac{\partial T}{\partial\phi_i}T^{-1}\dfrac{\partial T^{-1}}{\partial\phi_j}P = \\ \quad = 2\dfrac{\partial T}{\partial\phi_i}T^{-1}\dfrac{\partial T}{\partial\phi_j}T^{-1}T^{-1}P = 2\dfrac{\partial T}{\partial\phi_i}\dfrac{\partial T}{\partial\phi_j}T^{-3}P \end{cases} \tag{3-6}$$

A similar simplification occurs for the information matrix.

Holland (1973) suggests a method for showing that:

$$\frac{1}{2}\operatorname{trace}\left(V\frac{\partial^2 V^{-1}}{\partial\theta_\ell\theta_m}\right) = \operatorname{trace}\left[(T^{-1}P)\frac{\partial(P^{-1}T)}{\partial\theta_\ell}\left((T^{-1}P)\frac{\partial(P^{-1}T)}{\partial\theta_m}\right)^T\right]. \tag{3-7}$$

To simplify this computation we will write $V=QQ^T$, $V^{-1}=Q^{-T}Q^{-1}$, where $Q=P^{-1}T$ is column triangular (lemma 2 will be used in succeeding computations). It is easy to show that

$$\frac{\partial^2 V^{-1}}{\partial\theta_\ell\theta_m} = 2V^{-1}\frac{\partial V}{\partial\theta_\ell}V^{-1}\frac{\partial V}{\partial\theta_m}V^{-1} - V^{-1}\frac{\partial^2 V}{\partial\theta_\ell\theta_m}V^{-1}.$$

Hence $\frac{1}{2}\operatorname{trace}\left(V\frac{\partial^2 V^{-1}}{\partial\theta_\ell\theta_m}\right) = \operatorname{trace}\left(\frac{\partial V}{\partial\theta_\ell}V^{-1}\frac{\partial V}{\partial\theta_m}V^{-1}\right) - \frac{1}{2}\operatorname{trace}\left(\frac{\partial^2 V}{\partial\theta_\ell\theta_m}V^{-1}\right) =$

$$= \operatorname{trace}\left[\left(\frac{\partial Q}{\partial\theta_\ell}Q^T + Q\frac{\partial Q^T}{\partial\theta_\ell}\right)Q^{-T}Q^{-1}\left(\frac{\partial Q}{\partial\theta_m}Q^T + Q\frac{\partial Q^T}{\partial\theta_m}\right)Q^{-T}Q^{-1}\right] +$$

$$- \frac{1}{2}\operatorname{trace}\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\partial Q}{\partial\theta_m}Q^T + Q\frac{\partial Q^T}{\partial\theta_m}\right)Q^{-T}Q^{-1}\right] =$$

$$= \text{trace}\left[\left(\frac{\partial Q}{\partial \theta_\ell} Q^{-1} + \frac{\partial Q^T}{\partial \theta_\ell} Q^{-T}\right)\left(\frac{\partial Q}{\partial \theta_m} Q^{-1} + \frac{\partial Q^T}{\partial \theta_m} Q^{-T}\right)\right] +$$

$$- \frac{1}{2}\,\text{trace}\left[\left(\frac{\partial^2 Q}{\partial \theta_\ell \theta_m} Q^T + \frac{\partial Q}{\partial \theta_m}\frac{\partial Q^T}{\partial \theta_\ell} + \frac{\partial Q}{\partial \theta_\ell}\frac{\partial Q^T}{\partial \theta_m} + Q\frac{\partial^2 Q^T}{\partial \theta_\ell \theta_m}\right) Q^{-T} Q^{-1}\right].$$

Now Q is lower triangular, with 1 on it's main diagonal. So $\frac{\partial Q}{\partial \theta_\ell}$ is lower triangular with 0 on it's main diagonal. It follows that any product of lower triangular matrices involving $\frac{\partial Q}{\partial \theta_\ell}$ will have 0 trace. Similarly for $\frac{\partial Q}{\partial \theta_m}$ and for products of upper triangular matrices involving $\frac{\partial Q^T}{\partial \theta_\ell}$ and $\frac{\partial Q^T}{\partial \theta_m}$. Using this fact we obtain

$$\frac{1}{2}\,\text{trace}\left(V\frac{\partial^2 V^{-1}}{\partial \theta_\ell \theta_m}\right) = \text{trace}\left(\frac{\partial Q}{\partial \theta_\ell} Q^{-1}\frac{\partial Q^T}{\partial \theta_m} Q^{-T}\right) +$$

$$+ \text{trace}\left(\frac{\partial Q}{\partial \theta_m} Q^{-1}\frac{\partial Q^T}{\partial \theta_\ell} Q^{-T}\right) +$$

$$- \frac{1}{2}\,\text{trace}\left(\frac{\partial Q}{\partial \theta_m} Q^{-1}\frac{\partial Q^T}{\partial \theta_\ell} Q^{-T}\right) - \frac{1}{2}\,\text{trace}\left(\frac{\partial Q}{\partial \theta_\ell} Q^{-1}\frac{\partial Q}{\partial \theta_m} Q^{-T}\right).$$

But $\text{trace}(A^T) = \text{trace}(A)$, for any matrix A, so we finally get

$$\frac{1}{2}\,\text{trace}\left(V\frac{\partial^2 V^{-1}}{\partial \theta_\ell \theta_m}\right) = \text{trace}\left[\left(\frac{\partial Q}{\partial \theta_\ell} Q^{-1}\right)\left(\frac{\partial Q}{\partial \theta_m} Q^{-1}\right)^T\right] =$$

$$= \text{trace}\left[\left(Q^{-1}\frac{\partial Q}{\partial \theta_\ell}\right)\left(Q^{-1}\frac{\partial Q}{\partial \theta_m}\right)^T\right]$$

(because Q is column triangular).

Now having established 3-7 we note that

$$\frac{\partial}{\partial\theta_\ell}(P^{-1}T) = \frac{\partial P^{-1}}{\partial\theta_\ell}\ T + P^{-1}\frac{\partial}{\partial\theta_\ell}T =$$

$$= \quad -P^{-1}\frac{\partial P}{\partial\theta_\ell}\ P^{-1}T + P^{-1}\frac{\partial T}{\partial\theta_\ell} \quad .$$

So

$$\begin{cases} (T^{-1}P)\ \dfrac{\partial}{\partial\rho_\ell}(P^{-1}T) = -T^{-1}\dfrac{\partial P}{\partial\rho_\ell}\ P^{-1}T = -\dfrac{\partial P}{\partial\rho_\ell}\ P^{-1} \\ (T^{-1}P)\ \dfrac{\partial}{\partial\phi_\ell}(P^{-1}T) = T^{-1}\dfrac{\partial T}{\partial\phi_\ell} = \dfrac{\partial T}{\partial\phi_\ell}\ T^{-1} \end{cases} \tag{3-8}$$

Expressions (3-5), (3-6), (3-7) and (3-8), together with lemma 2, permit efficient computation of the expressions (2-8) required for the computation of Newton's step (2-7). First we notice that $\frac{\partial P}{\partial \rho_j} A$, and $\frac{\partial T}{\partial \rho_j}$ merely shift the columns of A down by j places, and append j zeros to the top of A. Furthermore, since Af is a vector if A is a matrix, and P and T are both column triangular, we see that we will never need to actually compute any nxn matrices (since we need only compute expressions of the form Av, where v is a vector, and A is column triangular, so this computation can be done trivially without expanding A into an nxn matrix). Specifically, we can break the computation down as follows:

Let $A = T^{-1}P$

1) Compute and store Af (requires n cells)

2) Compute and store Af' and A'f (requires n x (k+p) cells)

2.1) The gradient G is now given by computing $[Af']^T [Af]$ and $[A'f]^T [Af]$

2.2) Compute and store the "X^TX" matrix, that is

$$\begin{bmatrix} [Af']^T [Af'] & [Af']^T [A'f] \\ [Af']^T [A'f] & [A'f]^T [A'f] \end{bmatrix}$$

(requires (p+k) x (p+k) cells)

3) Compute and add to the matrix computed in 2.2 the nonlinear correction terms due to the second derivatives, that is

$$\begin{bmatrix} [Af'']^T [Af] & [A'f']^T [Af] \\ [A'f']^T [Af] & [A''f]^T [Af] \end{bmatrix}$$

(requires only n x 1 cells of temporary storage).

This gives us the Hessian H.

Of course, the computation of A' and A'' must be further broken down into special cases, depending on whether $A = T^{-1}P$ or just T^{-1} or P. This is done using specializations of (3-5) and (3-6). Note that in both steps 2 and 3 we have only had to compute products of the form Qv, where Q is column triangular, and v is a vector. As pointed out previously, this can easily be done given only the first column of Q. Steps 2.1, 2.2, and 3 also require the computation of inner products; again, this can be done easily with no need for additional storage. In fact, the entire algorithm given above does not use significantly more storage than that required for an ordinary regression, and, thanks to the special forms of the matrices involved, the required derivatives are computed fairly efficiently. A similar algorithm works for the information matrix $I(\gamma)$; the upper left k x k corner is given by $[Af']^T [Af']$, so we need only worry about the lower right p x p corner. Using (3-7) and part 4 of lemma 2, we see that it suffices to compute the p matrices given in formula (3-8), and then to compute the trace of all cross products. Again, note that column triangularity allows us to compute only the first columns of matrices, so we need only n cells for each matrix, rather than n^2.

In addition to the above, several interesting special results can be gleaned easily from the simplified forms (3-5), (3-6), (3-7), and (3-8). They are discussed in the next section.

IV. SPECIAL RESULTS

We specialize to $m(\beta) \equiv 0$, so we have a pure Box-Jenkins model, $f(\beta) = Y$.

i) If $T(\phi) = I$, so that we have a pure autoregressive model, then defining $Y_i = 0$ for $i \leq 0$ we have

$$G_j = \left[\frac{\partial P_Y}{\partial \rho_j}\right]^T PY$$

$$= \sum_{k=j+1}^{n} Y_{k-j} \left[Y_k - \sum_{i=1}^{p} \rho_i Y_{k-i}\right] \tag{4-1}$$

$$= \sum_{k=j+1}^{n} Y_{k-j} Y_k - \sum_{i=1}^{p} \rho_i \sum_{k=j+1}^{n} Y_{k-j} Y_{k-i} \quad .$$

Hence we want

$$\rho_1 \sum_{k=j+1}^{n} Y_{k-j} Y_{k-1} + \ldots + \rho_p \sum_{k=j+1}^{n} Y_{k-j} Y_{k-p} = \sum_{k=j+1}^{n} Y_{k-j} Y_k \,, \tag{4-2}$$

$$\text{for } j=1, \ldots, p \,.$$

Equation (4-1) is illuminating: it shows that (at least for this case) the second half of equations (2-5) reduce to an orthogonality condition: residuals orthogonal to the "data", where now the data is Y, rather than X as usual. This is reasonable because in this type of model, Y acts like X in the usual setup. In fact, recalling (1-1) and (3-1) our model is now

$$Y \sim N_n \left(0, \sigma^2 P^{-1}[P^{-1}]^T\right) \quad .$$

That is

$$PY \sim N_n\,(0,\ \sigma^2\, I),$$

or, expanding the product PY,

$$\begin{cases} Y_1 = \varepsilon_1 \\ Y_2 = \rho_1 Y_1 + \varepsilon_1 \\ \quad\vdots \\ Y_k = \rho_1 Y_{k-1} + \ldots + \rho_p Y_{k-p} + \varepsilon_k \\ \quad\vdots \\ Y_n = \rho_1 Y_{n-1} + \ldots + \rho_p Y_{n-p} + \varepsilon_n \end{cases}$$

where $\varepsilon \sim N_n\,(0,\ \sigma^2\, I)$

Clearly this is formally identical to the usual linear regression model $Y = X\beta+\varepsilon$, where here

$$X = \begin{pmatrix} 0 & 0 & & \\ Y_1 & 0 & & \\ Y_2 & Y_1 & \cdot\ \cdot\ \cdot & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ Y_{n-1} & Y_{n-2} & & \end{pmatrix}.$$

So indeed $[Y-X\beta]^T X = 0$ is equivalent to (4-1).

The usual method for estimating pure autoregressive models is to solve the Yule-Walker equations (see Box and Jenkins (1970) 3.2.2). In our notation these equations are:

$$\frac{\sum_{k=j+1}^{n}(Y_{k-j}-\bar{Y})(Y_k-\bar{Y})}{\sum_{k=1}^{n}(Y_k-\bar{Y})^2} = \rho_1 \frac{\sum_{k=j}^{n}(Y_{k-j+1}-\bar{Y})(Y_k-\bar{Y})}{\sum_{k=1}^{n}(Y_k-\bar{Y})^2} + \ldots$$

$$+ \rho_p \frac{\sum_{k=j+p-1}^{n}(Y_{k-j+p}-\bar{Y})(Y_k-\bar{Y})}{\sum_{k=1}^{n}(Y_k-\bar{Y})^2}, \quad j = 1, \ldots, p . \tag{4-3}$$

Eliminating the common term $\sum_{k=1}^{n}(Y_k-\bar{Y})^2$, and recalling that in (4-2) $Y_i=0$ for $i\leq 0$, we see that (4-3) differs from (4-2) only in that 0 is substituted for $\bar{Y}$. This makes sense, because the assumption $m(\beta)\equiv 0$ implies $EY=0$.

So we see that our method of estimation reduces to the usual one in this simple case. If we assume $m(\beta)=\beta$ (a scalar), so that $EY=\beta$, then the two estimation methods are similar, but not identical, since we estimate β simultaneously with ρ, rather than merely setting $\hat{\beta}=\bar{Y}$. (In practice, $\hat{\beta}$ usually is close to $\bar{Y}$).

ii) If $T(\phi) = I$ and $\rho = 0$, then

$$I(\theta) = \begin{pmatrix} n-1 & 0 & 0 & \cdots & 0 \\ 0 & n-2 & 0 & \cdots & 0 \\ 0 & 0 & n-3 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & n-p \end{pmatrix} .$$

This is a perfectly sensible answer, since we see from the above equations that the estimate for ρ_j is essentially based on n-j observations. In particular, for p=1

$$\hat{\rho} = \frac{\sum_{i=2}^{n} Y_i Y_{i-1}}{\sum_{i=2}^{n} Y_i^2} .$$

iii) Furthermore, from formulas (3-7) and (3-8) we see that if either $P(\rho) \equiv I$ or $T(\rho) \equiv I$, so that we have only ϕ's or only ρ's to estimate, the value of $I(\gamma)$ will depend only on the value of the ϕ or ρ vector, and not on whether or not it is a ϕ vector or a ρ vector. That is, $I(\phi)=I(\rho)$ whenever $\phi=\rho$ and, respectively, $P(\rho)=I$ or $T(\phi)=I$.

This result is rather surprising: it says that the asymptotic variance for the ρ's is the same as that for the ϕ's if only ρ's or ϕ's are present, even though they represent quite different models:

One is

$$Y_i - \rho_1 Y_{2-1} \cdots - \rho_p Y_{i-p} \sim N(0, \sigma^2)$$

The other is

$$Y_i \sim \varepsilon_i - \phi_1 \varepsilon_{i-1} \cdots - \phi_p \varepsilon_{i-p}$$

where

$$\varepsilon_1, \ldots, \varepsilon_n \sim \text{i.i.d. } N(0, \sigma^2)$$

iv) If $\rho=\phi$, then $I(\phi)$ is singular, since it has the form $\begin{pmatrix} A & -A \\ -A & A \end{pmatrix}$. This means that the parameters are not estimable, and this is reasonable since our model is now

$$Y \sim N(0, \sigma^2 I_n)$$

and many choices of ρ and ϕ will give us this model.

V. APPROXIMATIONS TO THE COVARIANCE MATRIX

We have assumed that $I^{-1}(\hat{\gamma}) \xrightarrow{P} I(\gamma)$, and in fact Rao (1965) shows that if F_n is the distribution function of γ and G_n is the distribution function of a random variable distributed $N(0, I^{-1}(\hat{\gamma}))$, then

$$\lim_{n\to\infty} |F_n - G_n| = 0 .$$

By the strong law of large numbers, and consistency we also have

$$\frac{1}{n} H(\hat{\gamma}) - \frac{1}{n} I(\gamma) \xrightarrow{P} 0, \quad \text{since } EH(\gamma) = E\, I(\gamma) .$$

(Note that it is <u>not</u> true that $H(\hat{\gamma}) \xrightarrow{P} I(\gamma)$, in fact $H(\gamma)$ need not converge to $I(\gamma)$, as we will see later.)

On the basis of this result, it has been suggested that we use $H(\hat{\gamma})$ rather than $I(\hat{\gamma})$ as an estimate of $I(\gamma)$. We point out some disadvantages to this approach.

i) Suppose that $f(\beta) = Y$, and that $\hat{\rho} = \hat{\phi}$, so that $I(\hat{\gamma})$ is singular. $H(\hat{\gamma})$ is not necessarily singular; in fact, let $\hat{\rho} = \hat{\phi} = 0$, and $p=2$. Then

$$H(\hat{\gamma}) = \begin{pmatrix} \sum_{i=2}^{n} Y_i^2 & \sum_{i=2}^{n} Y_i^2 - \sum_{i=3}^{n} Y_{i-2}Y_i \\ \sum_{i=2}^{n} Y_i^2 - \sum_{i=3}^{n} Y_{i-2}Y_i & \sum_{i=2}^{n} Y_i^2 + \sum_{i=3}^{n} Y_{i-2}Y_i \end{pmatrix} .$$

ii) Let $f(\beta) = Y-\beta$, $K = 1$, $T(\phi) = I$, $p = 1$, $\sigma = 1$.

Then

$$H(0) = \begin{pmatrix} n & Y_1-Y_n \\ Y_1-Y_n & \sum_{i=2}^{n} Y_{i-1}^2 \end{pmatrix}$$

Whereas

$$I(0) = \begin{pmatrix} n & 0 \\ 0 & n-1 \end{pmatrix}$$

The form for H(0) is most easily derived by observing that here

$$2 \log L(f, \beta, \rho) = (Y_1-\beta)^2 + \sum_{i=2}^{n} [(Y_i-\beta) - \rho(Y_{i-1}-\beta)]^2$$

So

$$\frac{\partial \log L}{\partial \rho} = - \sum_{2}^{n} [(Y_i-\beta) - \rho(Y_{i-1}-\beta)] (Y_{i-1}-\beta)$$

$$\frac{\partial \log L}{\partial \beta} = - (Y_i-\beta) + 2 \sum_{2}^{n} [(y_i-\beta) - \rho(Y_{i-1}-\beta)] [-1+\rho]$$

$$\frac{\partial \log L}{\partial \rho^2} = \sum_{2}^{n} (Y_{i-1}-\beta)^2$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = 1 + \sum_{2}^{n} [-1+\rho] [-1+\rho] = N - 2(N-1) \rho + (N-1) \rho^2$$

$$\frac{\partial^2 \log L}{\partial \rho \, \beta} = -\sum_{2}^{n} [-1+\rho] (Y_{i-1}-\beta) + \sum_{2}^{n} [(Y_i-\beta) - \rho(Y_{i-1}-\beta)] (-1)$$

$$= \sum_{2}^{n} (Y_{i-1}-\beta) - \rho \sum_{2}^{n} (Y_{i-1}-\beta) - \sum_{2}^{n} (Y_i-\beta) + \rho \sum_{2}^{n} (Y_{i-1}-\beta)$$

$$= (Y_1-\beta) - (Y_n-\beta) = Y_1-Y_n \cdot$$

Clearly H(0) does not converge in probability to I(0); however, under the assumption $\rho=0$

$$Y_1-Y_n \sim N(0,\ 2) \quad , \text{ and}$$

$$\sum_{i=2}^{n} Y_{i-1}^2 \sim X_{n-1}^2 \quad ,$$

so we see that

$$\frac{Y_1-Y_n}{n} = O_p(\frac{1}{n})$$

and

$$\frac{1}{n}\left(\sum_{i=2}^{n} Y_{i-1}^2 - n+1\right) = O_p(\frac{1}{n}) \quad .$$

In this case, however, I(0) is the correct answer, so we see that H(0) is not as good.

There is another approximation which is clearly superior to $H(\hat{\gamma})$:

$$H_2(\hat{\gamma}) = \begin{pmatrix} [f']^T V^{-1} f' & 0 \\ 0 & f^T[V^{-1}]'' f \end{pmatrix} \quad . \tag{5-1}$$

This is obtained by eliminating those components in (2-8) whose expectation is obviously 0. For the example we have

$$H_2(0) = \begin{pmatrix} n & 0 \\ 0 & \sum_{i-2}^{n} Y_{i-1}^2 \end{pmatrix},$$

which is still not as good as I(0). H_2 also still suffers from disadvantage i) above; in fact, the lower right corner of H_2 is identical to that of H.

We conclude that the variance of the θ's (Box-Jenkins' parameters) should not be estimated from $H(\hat{\gamma})$, but from $I(\hat{\gamma})$, since the two can differ significantly: a numerical example follows.

We generated Y by taking 100 points from a normal (0,1) distribution, so that $Y \sim N_n(0,I)$. Then we fit the model (1-1) with $m(\beta) = \beta$, where β is a scalar and $\theta = \begin{pmatrix}\rho\\ \phi\end{pmatrix}$, so that we fit a first order moving average, first order autoregressive process. (I.e., both P and T are present, but each depends only on one parameter.) We found

$$\hat{\gamma} = \begin{pmatrix} .13536 \\ -.36 \\ -.3125 \end{pmatrix},$$

$$H^{-1}(\hat{\gamma}) = \begin{pmatrix} .009328 & .000476 & .000826 \\ & .347967 & .345833 \\ & & .350404 \end{pmatrix},$$

$$I^{-1}(\hat{\gamma}) = \begin{pmatrix} .009319 & 0 & 0 \\ & 3.1134 & 3.16488 \\ & & 3.22633 \end{pmatrix}$$

Since admissibility requires $|\rho| \leq 1$, $|\phi| \leq 1$, this last expression means that ρ and φ are essentially inestimable.

It is to be noted that the large observed variances for $\hat{\rho}$ and

$\hat{\phi}$ are not accidental: if we had found $\hat{\rho} = -.36$, $\hat{\phi} = -.36$, then $I(\hat{\gamma})$ would have been singular, and the variances would have been infinite. In fact, if we fix $\hat{\rho}$ at -.36 and vary $\hat{\phi}$, we get a smooth progression from reasonable variance estimates to absurdly large ones.

$\hat{\phi}$	Estimated variance of $\hat{\phi}$
0	.0787
-.2	.33
-.3	2.06

One might conclude from this example that the estimated variances given by $H^{-1}(\hat{\gamma})$ are absurd.

In this context Wall (1973) has suggested looking at the estimated correlation matrix for ρ and ϕ, This is

$$\begin{pmatrix} 1 & .99041 \\ & 1 \end{pmatrix} \quad \text{for } H^{-1}(\hat{\gamma}) \text{ and}$$

$$\begin{pmatrix} 1 & .99859 \\ & 1 \end{pmatrix} \quad \text{for } I^{-1}(\hat{\gamma}) \ .$$

This indicates at once that the estimates for ρ and ϕ are unreliable, since they are so highly correlated. We could also look at the condition number for the covariance matrix of ρ and ϕ. For H^{-1} the eigenvalues are .0033505, .695021, the condition number 207; for I^{-1} .00448, 6.33525 and 1,414. The condition numbers for the correlation matrices are 208 for H^{-1} and 1,417 for I^{-1}. So we see that in fact the estimated covariance matrix is nearly singular, for H^{-1} as well as I^{-1}; this

indicates that the parameters are "nearly inestimable". That is, we can reasonably conjecture that the estimated variances given by H^{-1} are much too small.

This example points out that blind acceptance of variances estimated from H^{-1}, without examination of correlation coefficients, eigenvalues or condition numbers, can be quite misleading for this class of problems.

VI. NUMERICAL CONSIDERATIONS

Assuming that $m(\beta)$ is not very nonlinear, one would expect that if $T(\phi) = I$, Newton's method would work well, since then the model is almost linear: if $m(\beta)$ is linear the nonlinearity is caused by the presence of products $\rho_i\beta_j$. This is in fact the case. However, when $T(\phi)$ is present the model is strongly nonlinear, and, as one would expect, straight Newton's method does not work very well. Various schemes to insure convergence have been found to help: these are all based on the principle that the objective function $f^TV^{-1}f$ should not be allowed to increase from one iteration to the next. If the step based on Newton's method would cause an increase, it is not taken, but a step based on some sort of gradient method is taken instead. The specific algorithm that was found to be most effective is a derivative based modification of Powell's (1970) dog-leg, which was suggested by John Dennis.

Even with this method, however, we have encountered models where G was not zero, yet $H^{-1}G$ was. This means the algorithm got stuck in a valley or "rut", even though a minimum had not been found. The only way out would be to start again with a different initial guess.

A further problem for which we have no solution is that not all values of θ are allowable. The admissibility condition given by Box and Jenkins (1970, pp. 54 and 67) is rather messy to compute, so we do not attempt to verify admissibility of the estimated $\hat{\theta}$. As a consequence we may occasionally return ridiculous estimates. In general, as pointed out by Box and Jenkins, great care should be exercised when fitting this sort of model.

Finally a few words on initial guesses. The following seemed to work reasonably well.

1) Fit $f(\beta)$ by ordinary nonlinear least squares. Let $r = f(\hat{\beta})$.

2) Fit the Box-Jenkins model for $P(\rho)$ only to r. Let these new residuals be r_2.

3) Fit the Box-Jenkins model for $T(\phi)$ only to r_2.

4) Use the estimated β, ρ, ϕ as initial values for the full model.

REFERENCES

Box, G.E.P. and Jenkins, G.M. (1970), Time Series Analysis, Holden-Day, San Francisco, California.

Holland, P.W. (1973), personal communication.

Powell, M.J.O. (1970), A New Algorithm for Unconstrained Optimization, Nonlinear Programming (Rosen, Mangasarian, Ritter, editors), Academic Press.

Rao, C.R. (1965), Linear Statistical Inference and Its Applications, John Wiley and Sons, New York, New York.

Wall, K. (1973), personal communcation.