ASYMPTOTICALLY EFFICIENT ESTIMATION OF COVARIANCE MATRICES WITH LINEAR STRUCTURE[1]

By T. W. Anderson

Stanford University

One or more observations are made on a random vector, whose covariance matrix may be a linear combination of known symmetric matrices and whose mean vector may be a linear combination of known vectors; the coefficients of the linear combinations are unknown parameters to be estimated. Under the assumption of normality, equations are developed for the maximum likelihood estimates.

These equations may be solved by an iterative method; in each step a set of linear equations is solved. If consistent estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$ are used to obtain the coefficients of the linear equations, the solution of these equations is asymptotically efficient as the number of observations on the random vector tends to infinity. This result is a consequence of a theorem that the solution of the generalized least squares equations is asymptotically efficient if a consistent estimate of the covariance matrix is used. Applications are made to the components of variance model in the analysis of variance and the finite moving average model in time series analysis.

1. Introduction. This paper deals with estimation problems in which one or more observations are made on a p-component vector $\mathbf{X}$ with mean vector $\mathscr{E}\mathbf{X} = \boldsymbol{\mu}$ and covariance matrix $\mathscr{C}(\mathbf{X}) = \mathscr{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \boldsymbol{\Sigma}$. The mean vector may be a linear combination

$$(1) \qquad \boldsymbol{\mu} = \sum_{j=1}^{r} \beta_j \mathbf{z}_j$$

of known p-component vectors, $\mathbf{z}_1, \cdots, \mathbf{z}_r$, which are assumed (for convenience) to be linearly independent. The covariance matrix may be a linear combination

$$(2) \qquad \boldsymbol{\Sigma} = \sum_{g=0}^{m} \sigma_g \mathbf{G}_g$$

of known symmetric $p \times p$ matrices $\mathbf{G}_0, \mathbf{G}_1, \cdots, \mathbf{G}_m$, which are assumed to be linearly independent; it is also assumed that there is at least one set $\sigma_0, \sigma_1, \cdots, \sigma_m$ such that (2) is positive definite. We want to estimate the set on the basis of N observations $\mathbf{x}_1, \cdots, \mathbf{x}_N$.

If $\boldsymbol{\Sigma}$ is known or known to within a constant of proportionality, the best linear unbiased estimates or Markov estimates of $\beta_1, \cdots, \beta_r$ are the solutions to

135

the normal equations

(3) $$\sum_{i=1}^{r} \mathbf{z}_j' \mathbf{\Sigma}^{-1} \mathbf{z}_i \, \hat{\beta}_i = \mathbf{z}_j' \mathbf{\Sigma}^{-1} \bar{\mathbf{x}} \,, \qquad\qquad j = 1, \cdots, r \,,$$

where $\bar{\mathbf{x}}$ is the sample mean. If $\mathbf{X}$ has a normal distribution, (3) are the likelihood equations, obtained by setting equal to 0 the derivatives of the likelihood function with respect to $\beta_1, \cdots, \beta_r$, and the solution constitutes the maximum likelihood estimates. In any case the estimates are unbiased, $\mathscr{E}\hat{\beta}_i = \beta_i$, $i = 1, \cdots, r$, and the covariance matrix of the estimates is

(4) $$[\mathscr{C}(\hat{\beta}_i, \hat{\beta}_j)] = (1/N)[\mathbf{z}_i' \mathbf{\Sigma}^{-1} \mathbf{z}_j]^{-1} \,.$$

If μ is known, and $\mathbf{X}$ has a normal distribution, the maximum likelihood estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$ are a solution of the likelihood equations

(5) $$\operatorname{tr}\left(\sum_{h=0}^{m} \hat{\sigma}_h \, \mathbf{G}_h\right)^{-1} \mathbf{G}_g$$
$$= \operatorname{tr}\left(\sum_{h=0}^{m} \hat{\sigma}_h \, \mathbf{G}_h\right)^{-1} \mathbf{G}_g \left(\sum_{h=0}^{m} \hat{\sigma}_h \, \mathbf{G}_h\right)^{-1} \mathbf{C} \,, \qquad g = 0, 1, \cdots, m \,,$$

where

(6) $$\mathbf{C} = (1/N) \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})' \,;$$

these equations result from setting equal to 0 the derivatives of the likelihood function with respect to $\sigma_0, \sigma_1, \cdots, \sigma_m$. There is at least one solution $\hat{\sigma}_0, \hat{\sigma}_1, \cdots, \hat{\sigma}_m$ to (5) such that

(7) $$\hat{\mathbf{\Sigma}} = \sum_{g=0}^{m} \hat{\sigma}_g \, \mathbf{G}_g$$

is positive definite. If there is more than one solution to the likelihood equations, the absolute maximum to the likelihood function is attained by the solution minimizing $|\hat{\mathbf{\Sigma}}|$. The estimates are consistent and asymptotically efficient as $N \to \infty$; $N^{\frac{1}{2}}(\hat{\sigma}_0 - \sigma_0), N^{\frac{1}{2}}(\hat{\sigma}_1 - \sigma_1), \cdots, N^{\frac{1}{2}}(\hat{\sigma}_m - \sigma_m)$ have a limiting normal distribution with means 0 and covariance matrix

(8) $$[\tfrac{1}{2} \operatorname{tr} \mathbf{\Sigma}^{-1} \mathbf{G}_h \mathbf{\Sigma}^{-1} \mathbf{G}_g]^{-1} \,.$$

See Anderson[2] (1969), (1970).

If both μ and $\mathbf{\Sigma}$ are unknown and $\mathbf{X}$ is normally distributed, the likelihood equations are (3) and (5) with $\mathbf{\Sigma}$ replaced by (7) in (3) and μ replaced by $\hat{\mu} = \sum_{j=1}^{r} \hat{\beta}_j \mathbf{z}_j$ in (6). The estimates are consistent; the two sets are asymptotically independent; and each set of estimates has the same asymptotic distribution as when the other set of parameters is known.

The main purpose of this paper is to give a simple method of estimating $\sigma_0, \sigma_1, \cdots, \sigma_m$ that is asymptotically equivalent to maximum likelihood.

2. Estimation procedure. In view of (7) we can write (5) as

(9) $$\sum_{f=0}^{m} \operatorname{tr} \hat{\mathbf{\Sigma}}^{-1} \mathbf{G}_g \hat{\mathbf{\Sigma}}^{-1} \mathbf{G}_f \, \hat{\sigma}_f = \operatorname{tr} \hat{\mathbf{\Sigma}}^{-1} \mathbf{G}_g \hat{\mathbf{\Sigma}}^{-1} \mathbf{C} \,, \qquad g = 0, 1, \cdots, m \,.$$

These equations suggest an iterative procedure. Suppose μ is known. Let $\hat{\sigma}_0^{(0)}$, $\hat{\sigma}_1^{(0)}, \cdots, \hat{\sigma}_m^{(0)}$ be an initial set of values. Let $\hat{\sigma}_0^{(i)}, \hat{\sigma}_1^{(i)}, \cdots, \hat{\sigma}_m^{(i)}$ be the solutions

---

to the set of linear equations

(10) $\sum_{f=0}^{m} \operatorname{tr} \hat{\Sigma}_{i-1}^{-1} \mathbf{G}_g \hat{\Sigma}_{i-1}^{-1} \mathbf{G}_f \hat{\sigma}_f = \operatorname{tr} \hat{\Sigma}_{i-1}^{-1} \mathbf{G}_g \hat{\Sigma}_{i-1}^{-1} \mathbf{C}$, $g = 0, 1, \cdots, m$,

where

(11) $\hat{\Sigma}_{i-1} = \sum_{h=0}^{m} \hat{\sigma}_h^{(i-1)} \mathbf{G}_h$, $i = 1, 2, \cdots$.

If $\hat{\Sigma}_{i-1}$ is nonsingular, the matrix of coefficients in (10) is positive definite. [The proof of this statement was given by T. W. Anderson (1970) for $\hat{\Sigma}_{i-1}$ replaced by $\mathbf{I}$.] The iteration may be stopped at the ith stage if $\hat{\sigma}_0^{(i)}, \hat{\sigma}_1^{(i)}, \cdots, \hat{\sigma}_m^{(i)}$ do not differ by much from $\hat{\sigma}_0^{(i-1)}, \hat{\sigma}_1^{(i-1)}, \cdots, \hat{\sigma}_m^{(i-1)}$.

Since $\mathscr{E}\mathbf{C} = \Sigma$, given by (2), unbiased and consistent estimates of $\sigma_0, \sigma_1, \cdots$, σ_m can be obtained as the solution to

(12) $\sum_{f=0}^{m} \operatorname{tr} \Theta \mathbf{G}_g \Theta \mathbf{G}_f \hat{\sigma}_f = \operatorname{tr} \Theta \mathbf{G}_g \Theta \mathbf{C}$, $g = 0, 1, \cdots, m$,

for an arbitrary positive definite matrix Θ. [T. W. Anderson (1970) suggested $\Theta = \mathbf{I}$.] To obtain asymptotically efficient estimates only one step in the iteration is needed if the initial estimates are consistent (Theorem 2).

If $\beta_1, \cdots, \beta_r$ are to be estimated as well as $\sigma_0, \sigma_1, \cdots, \sigma_m$, we can obtain initial estimates $\hat{\beta}_1^{(0)}, \cdots, \hat{\beta}_r^{(0)}$ on the basis of (3) with Σ^{-1} replaced by Θ. Then define $\mathbf{C}_0$ by (6) with μ replaced by $\hat{\mu}^{(0)} = \sum_{j=1}^{r} \hat{\beta}_j^{(0)} \mathbf{z}_j$ and obtain initial estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$ from (12) with $\mathbf{C}$ replaced by $\mathbf{C}_0$. The iteration proceeds by using (3) with Σ replaced by $\hat{\Sigma}_{i-1}$ to obtain $\hat{\beta}_1^{(i)}, \cdots, \hat{\beta}_r^{(i)}$, then $\hat{\mu}^{(i)} = \sum_{j=1}^{r} \hat{\beta}_j^{(i)} \mathbf{z}_j$, then

(13) $\mathbf{C}_i = (1/N) \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})' + (\bar{\mathbf{x}} - \hat{\mu}^{(i)})(\bar{\mathbf{x}} - \hat{\mu}^{(i)})'$,

and then $\hat{\sigma}_0^{(i)}, \hat{\sigma}_1^{(i)}, \cdots, \hat{\sigma}_m^{(i)}$ by (10) with $\mathbf{C}$ replaced by $\mathbf{C}_i$. The matrix Θ can be $\mathbf{I}$ or a guess at Σ^{-1}.

The solution of (10) requires evaluation of quantities such as $\operatorname{tr} \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} \mathbf{L}$, where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{L}$ are symmetric and $\mathbf{A}$ is positive definite. Finding $\mathbf{A}^{-1}$ corresponds to solving $\mathbf{A}\mathbf{X} = \mathbf{I}$. The "forward solution" of a method of pivotal condensation or successive elimination corresponds to multiplying this equation on the left by a lower triangular matrix $\mathbf{F}$ to obtain $\mathbf{T}\mathbf{X} = \mathbf{F}$, where $\mathbf{T}$ is upper triangular. Then $\mathbf{F}\mathbf{A}\mathbf{F}' = \mathbf{T}\mathbf{F}'$ is diagonal and the diagonal elements are positive. Call the (positive diagonal) square root of this matrix $\mathbf{D}$, and let $\mathbf{H} = \mathbf{D}^{-1}\mathbf{F}$. Then $\mathbf{A}^{-1} = \mathbf{H}'\mathbf{H}$, and $\operatorname{tr} \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} \mathbf{L} = \operatorname{tr} \mathbf{H}'\mathbf{H}\mathbf{B}\mathbf{H}'\mathbf{H}\mathbf{L} = \operatorname{tr} \mathbf{H}\mathbf{B}\mathbf{H}'\mathbf{H}\mathbf{L}\mathbf{H}'$. Thus only the forward solution is needed. [See Section 2.3 of T. W. Anderson (1971 b) for more details.] In many applications the special forms of Σ make this easy to compute.

As pointed out in earlier papers, the problem is simplified if $\mathbf{G}_0 = \mathbf{I}$ and $\mathbf{G}_h = \mathbf{P}\Lambda_h \mathbf{P}'$, where Λ_h is diagonal, $h = 1, \cdots, m$, for some orthogonal matrix $\mathbf{P}$. Suppose

(14) $\Lambda_h = \begin{bmatrix} \nu_{1h}\mathbf{I} & 0 & \cdots & 0 \\ 0 & \nu_{2h}\mathbf{I} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \nu_{nh}\mathbf{I} \end{bmatrix}$, $h = 1, \cdots, m$,

where the orders of the Γ's are $p_1, p_2, \cdots, p_n$, respectively, not depending on h. Let $p_k V_k$ be the sum of the diagonal elements of $\mathbf{P'CP}$ in the position of ν_{kh} in $\mathbf{\Lambda}_h$. Then (5) is

$$(15) \qquad \sum_{k=1}^{n} \frac{p_k \nu_{kg}}{\sum_{h=0}^{m} \hat{\sigma}_h \nu_{kh}} = \sum_{k=1}^{n} \frac{p_k \nu_{kg} V_k}{\left(\sum_{h=0}^{m} \hat{\sigma}_h \nu_{kh}\right)^2}, \qquad g = 0, 1, \cdots, m,$$

and (10) is

$$(16) \qquad \sum_{f=0}^{m} \sum_{k=1}^{n} \frac{p_k \nu_{kg} \nu_{kf}}{\left(\sum_{h=0}^{m} \hat{\sigma}_h^{(i-1)} \nu_{kh}\right)^2} \hat{\sigma}_f^{(i)} = \sum_{k=1}^{n} \frac{p_k \nu_{kg} V_k}{\left(\sum_{h=0}^{m} \hat{\sigma}_h^{(i-1)} \nu_{kh}\right)^2},$$

$$g = 0, 1, \cdots, m.$$

3. Asymptotic efficiency. To obtain asymptotically efficient estimates as $N \to \infty$ we need only carry out the first stage of the iteration procedure. To show this we first consider the estimation of $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_r)'$ when $\boldsymbol{\Sigma}$ is estimated by a consistent estimate $\hat{\boldsymbol{\Sigma}}(N)$. Let the solution to (3) constitute $\hat{\boldsymbol{\beta}}(N)$, and let $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_r)$. Then $N^{\frac{1}{2}}[\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}]$ has a limiting normal distribution with covariance matrix $(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}$ even if $\mathbf{X}$ is not normal.

THEOREM 1. *Let* $\mathbf{x}_1, \cdots, \mathbf{x}_N$ *be identically independently distributed with mean* $\mathscr{E}\mathbf{X} = \mathbf{Z}\boldsymbol{\beta}$ *and covariance matrix* $\boldsymbol{\Sigma}$ *and let* $\hat{\boldsymbol{\Sigma}}(N)$ *be a consistent estimate of* $\boldsymbol{\Sigma}$. *Then, if* $\hat{\boldsymbol{\beta}}^*(N)$ *is the solution to* (3) *with* $\boldsymbol{\Sigma}$ *replaced by* $\hat{\boldsymbol{\Sigma}}(N)$, $N^{\frac{1}{2}}[\hat{\boldsymbol{\beta}}^*(N) - \boldsymbol{\beta}]$ *has a limiting normal distribution with covariance matrix* $(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}$. *If* $\hat{\boldsymbol{\beta}}(N)$ *is asymptotically efficient, so is* $\hat{\boldsymbol{\beta}}^*(N)$.

PROOF.

$$N^{\frac{1}{2}}[\hat{\boldsymbol{\beta}}^*(N) - \hat{\boldsymbol{\beta}}(N)]$$

$$(17) \qquad = N^{\frac{1}{2}}\{\hat{\boldsymbol{\beta}}^*(N) - \boldsymbol{\beta} - [\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}]\}$$

$$= \{[\mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}(N)\mathbf{Z}]^{-1}\mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}(N) - [\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\}N^{\frac{1}{2}}[\bar{\mathbf{x}}(N) - \mathbf{Z}\boldsymbol{\beta}]$$

converges stochastically to $\mathbf{0}$, where $\bar{\mathbf{x}}(N)$ is the mean of N observations, because

$$(18) \qquad \mathrm{plim}_{N\to\infty} [\mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}(N)\mathbf{Z}]^{-1}\mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}(N) = (\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}$$

and $N^{\frac{1}{2}}[\bar{\mathbf{x}}(N) - \mathbf{Z}\boldsymbol{\beta}] = N^{\frac{1}{2}}[\bar{\mathbf{x}}(N) - \boldsymbol{\mu}]$ has a limiting distribution. Thus $N^{\frac{1}{2}}[\hat{\boldsymbol{\beta}}^*(N) - \boldsymbol{\beta}]$ has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix $(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}$, which is the same as the limiting normal distribution of $N^{\frac{1}{2}}[-\hat{\boldsymbol{\beta}}(N)]$. If $\hat{\boldsymbol{\beta}}(N)$ is asymptotically efficient, then $\hat{\boldsymbol{\beta}}^*(N)$ is (in the same sense). When $\hat{\boldsymbol{\beta}}(N)$ is maximum likelihood, as when $\mathbf{X}$ has a normal distribution, it is asymptotically efficient in the sense of attaining the Cramér–Rao lower bound for the covariance matrix of unbiased estimates.

This theorem includes the case where $\boldsymbol{\Sigma}$ is a continuous function of a vector of parameters $\boldsymbol{\theta}$. Then $\boldsymbol{\Sigma}(N)$ may be the function of a consistent estimate $\hat{\boldsymbol{\theta}}(N)$ of $\boldsymbol{\theta}$. Note that $\hat{\boldsymbol{\theta}}(N)$ does not need to be independent of $\bar{\mathbf{x}}(N)$ and $N^{\frac{1}{2}}[\hat{\boldsymbol{\theta}}(N) - \boldsymbol{\theta}]$ does not need to be bounded in probability (as is usually required for this treatment of maximum likelihood estimates).

We shall now apply Theorem 1 to the estimation of $\sigma_0, \sigma_1 \cdots, \sigma_m$ when μ is known. Let $\mathbf{C} = (c_{ij})$, $\mathbf{\Sigma} = (\sigma_{ij})$, and $\mathbf{G}_h = (g_{ij}^{(h)})$. Let $\mathbf{c}$ be the vector with components c_{ij}, $i \leq j$. Let $\boldsymbol{\sigma}$ be the vector with components σ_{ij}, $i \leq j$, and $\mathbf{g}_h$ be the vector with components $g_{ij}^{(h)}$, $i \leq j$. Then

$$(19) \qquad \mathscr{E}\mathbf{c} = \boldsymbol{\sigma} = \sum_{h=0}^{m} \sigma_h \mathbf{g}_h ,$$

and under normality of $\mathbf{X}$ the covariance matrix of $\mathbf{c}$ is $\mathbf{\Phi} = (\phi_{ij,kl})$, where $\phi_{ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$, $i \leq j$, $k \leq l$. [See T. W. Anderson (1958), Section 4.2.3, for example.] Since [T. W. Anderson (1969)]

$$(20) \qquad \mathbf{g}_h' \mathbf{\Phi}^{-1} \mathbf{c} \equiv \tfrac{1}{2} \operatorname{tr} \mathbf{\Sigma}^{-1} \mathbf{G}_h \mathbf{\Sigma}^{-1} \mathbf{C} ,$$

(10) for $i = 1$ is identical to (3) with the $\mathbf{z}_j$'s replaced by $\mathbf{g}_h$'s, $\bar{\mathbf{x}}$ by $\mathbf{c}$, $\hat{\beta}_i$'s by $\hat{\sigma}_f$'s, and $\mathbf{\Sigma}$ by $\hat{\mathbf{\Phi}}_0$, a consistent estimate of $\mathbf{\Phi}$. Theorem 1 can be applied.

THEOREM 2. *Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be N observations from $N(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu}$ is known and $\mathbf{\Sigma}$ is given by (2). Let $\mathbf{C}$ be defined by (6). Let $\hat{\sigma}_0^{(0)}, \hat{\sigma}_1^{(0)}, \cdots, \hat{\sigma}_m^{(0)}$ be a set of consistent estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$. Let $\hat{\sigma}_0^{(1)}, \hat{\sigma}_1^{(1)}, \cdots, \hat{\sigma}_m^{(1)}$ be the solution to (10) for $i = 1$. Then $N^{\frac{1}{2}}(\hat{\sigma}_0^{(1)} - \sigma_0), N^{\frac{1}{2}}(\hat{\sigma}_1^{(1)} - \sigma_1), \cdots, N^{\frac{1}{2}}(\hat{\sigma}_m^{(1)} - \sigma_m)$ have a limiting normal distribution with means 0 and covariance matrix (8), and $\hat{\sigma}_0^{(1)}, \hat{\sigma}_1^{(1)}, \cdots, \hat{\sigma}_m^{(1)}$ are asymptotically efficient.*

If both sets $\beta_1, \cdots, \beta_r$ and $\sigma_0, \sigma_1, \cdots, \sigma_m$ are estimated, initial estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$ can be obtained from (12) with μ replaced by $\bar{\mathbf{x}}$ in (6). Then $\hat{\beta}_1^{(1)}, \cdots, \hat{\beta}_r^{(1)}$ and $\hat{\sigma}_0^{(1)}, \hat{\sigma}_1^{(1)}, \cdots, \hat{\sigma}_m^{(1)}$ satisfy Theorems 1 and 2 and are asymptotically efficient.

4. Applications. A general model for the analysis of variance can be written

$$(21) \qquad \mathbf{X} = \mathbf{Z}\boldsymbol{\beta} + \sum_{h=1}^{m} \mathbf{U}_h \mathbf{b}_h + \mathbf{e} ,$$

where $\mathbf{U}_h$ is a $p \times n_h$ matrix and $\mathbf{b}_h$ has the distribution $N(\mathbf{0}, \sigma_h \mathbf{I})$, $h = 1, \cdots, m$, $\mathbf{e}$ has the distribution $N(\mathbf{0}, \sigma_0 \mathbf{I})$, and $\mathbf{b}_1, \cdots, \mathbf{b}_m, \mathbf{e}$ are independent. Then $\mathbf{G}_0 = \mathbf{I}$ and $\mathbf{G}_h = \mathbf{U}_h \mathbf{U}_h'$. Hartley and J. N. K. Rao (1967) derived the likelihood equation for this model with one observation on $\mathbf{X}$. The two methods they propose for solving these equations are numerically more complicated than the method proposed in this paper.

C. R. Rao (1972) has considered (21) for $N = 1$. His method involves finding $\hat{\beta}_1^{(0)}, \cdots, \hat{\beta}_r^{(0)}$ as outlined in Section 2 of this paper, then $\hat{\mu}^{(0)}$ and $\mathbf{C}_0$, and using (12) for estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$.

The moving average stationary stochastic process of finite order is defined by $x_t = \mu + \sum_{g=0}^{m} \alpha_g v_{t-g}$, where $\mathscr{E}v_t = 0$, $\mathscr{E}v_t^2 = 1$, and $\mathscr{E}v_t v_s = 0$, $t \neq s$. Then $\mathscr{E}x_t = \mu$ and $\sigma_h = \mathscr{E}(x_t - \mu)(x_{t+h} - \mu) = \sum_{g=0}^{m-h} \alpha_g \alpha_{g+h}$, $h = 0, 1, \cdots, m$, and $\mathscr{E}(x_t - \mu)(x_{t+h} - \mu) = 0$ for $h > m$. Here $\mathbf{G}_0 = \mathbf{I}$ and $g_{st}^{(h)} = 1$, $|s - t| = h$, $g_{st}^{(h)} = 0$, $|s - t| \neq h$, $h = 1, \cdots, m$. If $N = 1$, the right-hand side of (10) is

$$(22) \qquad \operatorname{tr} \hat{\mathbf{\Sigma}}_{i-1}^{-1} \mathbf{G}_g \hat{\mathbf{\Sigma}}_{i-1}^{-1} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' = (\mathbf{x} - \boldsymbol{\mu})' \hat{\mathbf{\Sigma}}_{i-1}^{-1} \mathbf{G}_g \hat{\mathbf{\Sigma}}_{i-1}^{-1} (\mathbf{x} - \boldsymbol{\mu}) ,$$

where $\boldsymbol{\mu} = (\mu, \cdots, \mu)'$, which is a quadratic form in $\mathbf{y}$, the solution to $\hat{\boldsymbol{\Sigma}}_{i-1}\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$. The matrix $\hat{\boldsymbol{\Sigma}}_{i-1}$ has nonzero elements only on the main diagonal and on diagonals within m of the main diagonal; each row of $\hat{\boldsymbol{\Sigma}}_{i-1}$ has at most $2m + 1$ nonzero elements. In the forward solution by successive elimination of variables, for example, the elimination of one variable from all equations requires at most operations on $m + 1$ equations. At the end of the forward solution each equation has at most $m + 1$ unknowns. [See T. W. Anderson (1971b) for details.] The number of arithmetic operations to find $\mathbf{y}$ is only a multiple of mp, rather than multiple of p^2 as is the case with an arbitrary matrix of coefficients. The coefficients on the left-hand side of (10) can be approximated by using the fact that the inverse of the covariance matrix of a moving average process is approximately the covariance matrix of the autoregressive process with the same coefficients. [For example, see Anderson (1971a) and Shaman (1969).] The sample serial covariances may serve as initial estimates of $\sigma_0, \sigma_1, \cdots, \sigma_m$.

In this model one observation vector is a univariate time series of length p. Then the asymptotic theory as $p \to \infty$ may be useful. Whittle (1953), (1954), A. M. Walker (1964), and Hannan (1970) have treated maximum likelihood estimates in the context of a spectral density depending on a finite number of parameters. It follows that $p^{\frac{1}{2}}(\hat{\sigma}_0 - \sigma_0), p^{\frac{1}{2}}(\hat{\sigma}_1 - \sigma_1), \cdots, p^{\frac{1}{2}}(\hat{\sigma}_m - \sigma_m)$ have a limiting normal distribution with means 0 and a covariance matrix which is the inverse of the matrix with elements $\frac{1}{2}$ times

$$(23) \qquad \lim_{p \to \infty} p^{-1} \operatorname{tr} \boldsymbol{\Sigma}^{-1}\mathbf{G}_h \boldsymbol{\Sigma}^{-1}\mathbf{G}_g = 4(2\pi)^{-3}\delta_h\,\delta_g \int_{-\pi}^{\pi} \cos \lambda h \cos \lambda g f^{-2}(\lambda)\, d\lambda \,,$$

where $\delta_0 = \frac{1}{2}$ and $\delta_h = 1, h > 0$, and $f(\lambda)$ is the spectral density. [T. W. Anderson (1971b) has proved (23) directly; the result is valid for any continuous positive spectral density.]

Durbin (1959), A. M. Walker (1962), Hannan (1969), (1970), Box and Jenkins (1970), and Clevenson (1969) have given other methods for estimating the σ_h's or α_h's for the moving average process. The methods of Durbin and Walker are not asymptotically efficient; Box and Jenkins evaluate or approximate the likelihood; Hannan and Clevenson use the sample spectral density, involving a number of arithmetic operations proportional to $p \log p$.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] ANDERSON, T. W. (1969). Statistical inference for covariance matrices with linear structure, in *Proceedings of the Second International Symposium on Multivariate Analysis*, ed. P. R. Krishnaiah. Academic Press, New York, 55–66.

[3] ANDERSON, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices, in *Essays in Probability and Statistics*, ed. R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith. University of North Carolina Press, Chapel Hill, 1–24.

[4] ANDERSON, T. W. (1971a). *The Statistical Analysis of Time Series*. Wiley, New York.

[5] ANDERSON. T. W. (1971b). Estimation of covariance matrices with linear structure and moving average processes of finite order. Technical Report No. 6, Stanford Univ.

[6] Box, George E. P. and Jenkins, Gwilym M. (1970). *Time Series Analysis Forecasting and Control*. Holden-Day, San Francisco.

[7] Clevenson, M. Lawrence (1970). Asymptotically efficient estimates of the parameters of a moving avarage time series. Technical Report No. 15, Stanford Univ.

[8] Durbin, J. (1959). Efficient estimation of parameters in moving-average models. *J. Roy. Statist. Soc Ser. B* 22 139–153.

[9] Hannan, E. J. (1969). The estimation of mixed moving average autoregressive systems. *Biometrika* 56 579–593.

[10] Hannan, E. J. (1970). *Multiple Time Series*. Wiley, New York.

[11] Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* 54 93–108.

[12] Rao, C. R. (1972). Estimating variance and covariance components in linear models. *J. Amer. Statist. Assoc.* 67 112–115.

[13] Shaman, P. (1969). On the inverse of the covariance matrix of a first order moving average. *Biometrika* 56 595–608.

[14] Walker, A. M. (1962). Large sample estimation of parameters for autoregressive processes with moving-average residuals. *Biometrika* 49 117–131.

[15] Walker, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *J. Austral. Math. Soc.* 4 363–384.

[16] Whittle, P. (1953). Estimation and information in stationary time series. *Ark. Mat. Astr. Fys.* 2 423–434.

[17] Whittle, P. (1954). Appendix 2 to *A Study in the Analysis of Stationary Time Series* by H. Wold. Almqvist and Wiksell, Uppsala.

Department of Statistics
Stanford University
Stanford, California 94305