N O T I C E

THIS DOCUMENT HAS BEEN REPRODUCED FROM MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED IN THE INTEREST OF MAKING AVAILABLE AS MUCH INFORMATION AS POSSIBLE

A DISCRIMINANT APPROACH TO PARAMETER ESTIMATION IN THE LINEAR MODEL WITH UNKNOWN VARIANCE–COVARIANCE MATRIX

C. R. Hallum

National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas

and M. D. Pore

Lockheed Electronics Company, Inc.
Systems and Services Division
Houston, Texas

(E80-10192) A DISCRIMINANT APPROACH TO PARAMETER ESTIMATION IN THE LINEAR MODEL WITH UNKNOWN VARIANCE-COVARIANCE MATRIX (Lockheed Electronics Co.) 10 p HC A02/MF A01 CSCL 05B G3/43

N80-28789

Unclas 00192

Annual Meeting of the American Statistical Association

Chicago, Illinois

August 15–18, 1977

JSC–13036
LEC–10532

A DISCRIMINANT APPROACH TO PARAMETER ESTIMATION IN THE LINEAR MODEL WITH UNKNOWN VARIANCE-COVARIANCE MATRIX

C. R. Hallum
National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas

M. D. Pore
Lockheed Electronics Company, Inc.
Systems and Services Division
Houston, Texas

ABSTRACT

An estimate of the nonrandom vector, β, of parameters is obtained in the linear model $Y = X\beta + \varepsilon$, where ε is an unobservable random vector of disturbances and is assumed to satisfy $E(\varepsilon) = 0$ (the zero vector) and $E(\varepsilon\varepsilon^T) = V$, with V assumed unknown. The estimate obtained is the one which yields maximal similarity to the sample $Y_1, Y_2, \cdots, Y_N$ via the Sebestyen similarity function. Under the normality assumption, the resulting estimate is seen to be an unbiased estimate and justification given for selecting the maximum likelihood estimate for V in the Gauss-Markov estimate for B.

1. INTRODUCTION

Consider the linear model $Y = X\beta + \varepsilon$, where Y is an $n \times 1$ observable random vector, X is an $n \times m$ matrix of fixed elements and $\text{rank}(X) = m \leq n$, β is an $m \times 1$ nonrandom vector of parameters to be estimated and ε is an unobservable random vector of disturbances with ε assumed to satisfy $E(\varepsilon) = 0$ with unknown variance-covariance matrix $E(\varepsilon\varepsilon^T) = V$. It is well known that in case V is known (up to at least a scalar multiple), the Gauss-Markov theorem [1] applies and the best linear unbiased estimate of β is given by

$$\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y \qquad (1)$$

Other authors [2, 3, 4] have considered the problem of obtaining optimal estimates for β when V is unknown. Rao [4] showed that the estimate of β obtained by merely substituting an estimate $\hat{V}$ for V in equation (1) is not necessarily best; in particular, it may be possible to use known or inferred knowledge of the covariance V to obtain an estimator with better characteristics. Born [2] has written a recursive estimator when V is not known but is assumed to be block diagonal with equal diagonal blocks. McElroy [3] obtained necessary and sufficient conditions on V for equation (1) to be equivalent to the least-squares solution

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \qquad (2)$$

In this paper, we assume that the only available information is that contained in a sample $Y_1, Y_2, \cdots, Y_N$, and an estimate, $\hat{\beta}$, of β is obtained which results in maximal similarity to the given sample via the Sebestyen [5] similarity function. The resulting estimate appears in the form of equation (1), with V replaced by the standard (and in the normal theory case, the maximum likelihood) estimate of the variance-covariance matrix.

2. THE SEBESTYEN INTERSET SIMILARITY FUNCTION

If R^n denotes Euclidean n-space and P is the class of finite sequences of sample observations in R^n (i.e., $W, Z \in P$, provided $W = \{W_1, W_2, \cdots, W_N\}$ and $Z = \{Z_1, Z_2, \cdots, Z_M\}$ where $W_i, Z_j \in R^n$ for $i = 1, 2, \cdots, N$; $j = 1, 2, \cdots, M$), the Sebestyen [5] similarity function is defined as follows.

Definition: If $W, Z \in P$ with N and M elements, respectively, and if A is any $m \times n$ matrix, define the function $S_A: P \times P \to R_0$, where R_0 is the set of nonnegative real numbers, by

$$S_A(W,Z) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (W_i - Z_j)^T A^T A (W_i - Z_j) \qquad (3)$$

(The superscript T denotes the transpose.) Given a transformation A, $S_A(W,Z)$ is a measure of the similarity of the two samples W and Z in the transformed space (i.e., the resulting space after transforming R^n by A), and if W and Z are random samples from populations π_1 and π_2, respectively, then $S_A(W,Z)$ may be considered as a measure of the similarity of π_1 to π_2.

If $W,Z \in P$ have sample variance-covariance matrices $\hat{V}_1$ and $\hat{V}_2$, respectively, that is,

$$\left.\begin{aligned} \hat{V}_1 &= \frac{1}{N} \sum_{i=1}^{N} \left(W_i - \overline{W}\right)\left(W_i - \overline{W}\right)^T \\ \hat{V}_2 &= \frac{1}{M} \sum_{i=1}^{M} \left(Z_i - \overline{Z}\right)\left(Z_i - \overline{Z}\right)^T \end{aligned}\right\} \qquad (4)$$

and Tr denotes the trace operator [and $S(W,Z) \triangleq S_I(W,Z)$], then the properties below are easily verified.

Properties:

1. $S_A(W,Z) = \mathrm{Tr}\left[A(\hat{V}_1 + \hat{V}_2)A^T\right] + (\overline{W} - \overline{Z})^T A^T A (\overline{W} - \overline{Z})$.
2. $S(W,W) = 2\mathrm{Tr}(\hat{V}_1)$ and $S_A(W,W) = 2\mathrm{Tr}\left(A\hat{V}_1A^T\right)$.
3. $S_A(W,Z) = S_A(Z,W)$.
4. $S_A(W,Z) \geq 0$ for every $W,Z \in P$ and for each $m \times n$ matrix.
5. If $V \in R^n$ then $S_A(W,V) = S_A(W,\{V\}) = \mathrm{Tr}\left(A\hat{V}_1A^T\right)$ $+ (\overline{W} - V)^T A^T A (\overline{W} - V)$.
6. If $W = \{W_1, W_2, \cdots, W_N\}$ and $Z = \{Z_1, Z_2, \cdots, Z_M\}$, then $\frac{1}{M} \sum_{j=1}^{M} S_A(W,\{Z_j\}) = S_A(W,Z)$.

The Sebestyen decision rule is to classify an unknown u as belonging to category W provided

$$S_A(W,\{u\}) < S_B(Z,\{u\}) \tag{5}$$

where A and B are preselected transformations for W and Z, respectively. Consequently, the function $f(u) = S_B(Z,\{u\}) - S_A(W,\{u\})$ is the discriminant function for the Sebestyen decision rule, with classification of u into W or Z being accomplished by noting the sign of f(u); that is, u is classified as belonging to W or Z depending on whether $f(u) > 0$ or $f(u) < 0$, respectively.

3. A TRANSFORMATION TO MINIMIZE THE INTRASET DISTANCE

Thus far, no specifications have been placed on the transformation A; however, if A is an orthogonal matrix, the transformation results in a rotation of the original space whereby distances and, hence, angles are preserved. If the determinant of A [Det(A)] is 1, A is a volume-preserving transformation. The transformation of interest in this paper is specified in Theorem 1, the proof of which is dependent on the following well-known relationship between the arithmetic and geometric mean.

Lemma 1: If $d_i \geq 0$ for $i = 1, 2, \cdots, n$, then

$$\frac{1}{n}\sum_{i=1}^{n} d_i \geq \left(\prod_{i=1}^{n} d_i\right)^{1/n} \tag{6}$$

with equality holding if and only if $d_1 = d_2 = \cdots = d_n$.

Theorem 1: Under the condition Det(A) = 1 and S_A is positive definite, an $n \times n$ matrix A minimizes $S_A(W,W)$ [that is, the similarity of a set with itself] if and only if $A\hat{V}_1A^T = \lambda I$, where $\lambda = [\text{Det}(\hat{V}_1)]^{1/n}$ and $\hat{V}_1$ is the sample variance-covariance matrix of W specified in equation (4).

Proof: If B is any n × n matrix with Det(B) = 1 and A is the matrix specified in the hypothesis, from Property 3 and the fact that $A\hat{V}_1A^T = \lambda I$, we have $S_B(W,W) - S_A(W,W) = 2\text{Tr}\left(B\hat{V}_1B^T\right) - 2n\lambda$. Letting U be the orthogonal matrix such that $UB\hat{V}_1B^TU^T = D$, where D is diagonal, then

$$2\text{Tr}\left(B\hat{V}_1B^T\right) - 2n\lambda = 2\text{Tr}\left(UB\hat{V}_1B^TU^T\right) - 2n\lambda$$

$$= 2\text{Tr}(D) - 2n\lambda$$

$$= 2n\left[\frac{1}{n}\sum_{i=1}^{n} d_i - \left(\prod_{i=1}^{n} d_i\right)^{1/n}\right] \tag{7}$$

But equation (7) is nonnegative, by Lemma 1, with equality holding if and only if $d_1 = d_2 = \cdots = d_n$, in which case $B\hat{V}_1B^T = \lambda I$, which was to be demonstrated. Note that Aw exists if $\hat{V}_1$ is positive definite. Indeed Aw = EV, where E is the matrix whose columns are the eigenvectors of $\hat{V}_1$ and V is a diagonal matrix whose *ith* diagonal element is λ/β_i, i = 1, 2, ⋯, n, where β_i = *ith* eigenvalue of $\hat{V}_1$.

Theorem 1 associates with each sample W ∈ P, a transformation, A_W, with the property that A_W causes W to cluster in a spherical fashion after transformation with uncorrelated variates having equal variances. If W and Z are samples from populations π_1 and π_2 and if A and B are selected such that

$$A = \lambda_1^{-1/2}A_0 \quad ; \quad B = \lambda_2^{-1/2}B_0 \tag{8}$$

where $\lambda_1 = [\mathrm{Det}(\hat{V}_1)]^{1/n}$, $\lambda_2 = [\mathrm{Det}(\hat{V}_2)]^{1/n}$, and A_0 and B_0 are determined independently for W and Z, respectively, by Theorem 1, the effect is that of a normalization of the intraset similarity in that not only are the intraset distances minimal but $S_A(W,W) = S_B(Z,Z) = 2n$ as well (i.e., the normalization gives each intraset similarity the same value). Moreover, if instead of a threshold of zero in the Sebestyen decision rule [see eq. (5)], we choose the threshold

$$T = \ln \frac{\mathrm{Det}(\hat{V}_1) p_2}{\mathrm{Det}(\hat{V}_2) p_1} \tag{9}$$

the resulting decision rule is the Bayes maximum likelihood decision rule [6] (when the population distributions are Normal) with equal costs of misclassification and *a priori* probabilities p_1 and p_2 for π_1 and π_2, respectively.

4. THE ESTIMATE FOR β

In the weighted least-squares procedure, an estimator of β was selected which minimized $(Y - X\beta)^T V^{-1} (Y - X\beta)$. The optimal estimate is specified in equation (1) and the significance of such an estimator is that of being able to predict, or adjust in some applications, Y to a given matrix X. Since V is ordinarily unknown, we proceed as described below.

Collect N sample values; denote this sample by

$W = \{Y_1, Y_2, \cdots, Y_N\}$ and the sample variance-covariance matrix by

$$\hat{V} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)^T \tag{10}$$

If A is selected such that

$$A\hat{V}A^T = \lambda I \tag{11}$$

where

$$\lambda = [\mathrm{Det}(\hat{V})]^{1/n} \tag{12}$$

then Theorem 1 guarantees that the similarity of W with itself is a minimum after transformation by A. For prediction or adjustment purposes, what we now want to do is to select the vector $Z = X\beta$ which, after transformation, is more similar to the representative sample W than any other such vector. Equivalently, we want to select β such that $S_A(W,\{Z\})$ is a minimum where $Z = X\beta$ and A satisfies equation (11).

Theorem 2: The value of β which minimizes $S_A(W,X\beta)$ is given by

$$\hat{\beta} = \left(X^T\hat{V}^{-1}X\right)^{-1}X^T\hat{V}^{-1}\overline{Y} \tag{13}$$

where $\hat{V}$ is the sample variance-covariance matrix of W, A is the transformation specified in Theorem 1, and $\overline{Y}$ is the sample mean of W.

Proof: From Property 6 of S_A,

$$S_A^{\cdot}(W,X\beta) = \mathrm{Tr}(A\hat{V}A^T) + (\overline{Y} - X\beta)^T A^T A(\overline{Y} - X\beta) \tag{14}$$

Differentiating this expression with respect to β, equating to 0, and solving for β yields

$$\hat{\beta} = \left(X^TA^TAX\right)^{-1}X^TA^TA\overline{Y} \tag{15}$$

However, from the condition that $A\hat{V}A^T = \lambda I$

$$A^TA = (1/\lambda)\hat{V}^{-1} \tag{16}$$

which results in equation (8) after substitution into equation (9), which was to be demonstrated.

Under the normality assumption on Y [i.e., $y \sim MVN(X\beta, V)$] where V is unknown, $\hat{V}$ and $\bar{Y}$ are independent; therefore

$$\left.\begin{aligned} E(\hat{\beta}) &= E\left[\left(X^T\hat{V}^{-1}X\right)^{-1}X^T\hat{V}^{-1}\bar{Y}\right] \\ &= E\left[\left(X^T\hat{V}^{-1}X\right)^{-1}X^T\hat{V}^{-1}\right]E(\bar{Y}) \\ &= E\left[\left(X^T\hat{V}^{-1}X\right)^{-1}X^T\hat{V}^{-1}\right]X\beta \\ &= E\left[\left(X^T\hat{V}^{-1}X\right)^{-1}X^T\hat{V}^{-1}X\right]\beta \end{aligned}\right\} \qquad (17)$$

$$= \beta$$

Consequently, $\hat{\beta}$ is unbiased under these conditions.

5. SUMMARY

An estimate, $\hat{\beta}$, of β in the linear model $Y = X\beta + \varepsilon$ was obtained such that $X\hat{\beta}$ yielded maximal similarity to the sample Y_1, Y_2, $\cdots$, Y_N via the Sebestyen similarity function. The unobservable random error term, ε, was assumed to satisfy $E(\varepsilon) = 0$ (the zero vector) and $E(\varepsilon\varepsilon^T) = V$, with V assumed to be unknown. The resulting estimate is seen to be in the same form as the standard Gauss-Markov estimate

$$\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y \qquad (18)$$

except V is replaced by the standard (and under the normality assumption, the maximum likelihood) estimate of the variance-covariance matrix.

6. REFERENCES

1. Graybill, Franklin A.: Theory and Application of the Linear Model. Duxbury Press (Belmont, Calif.), 1976.

2. Born, H. G.; and Tapley, B. D.: Sequential Estimation of State and the Observation Error Covariance Matrix. Instit. Aero. Astro., 1969.

3. McElroy, F. M.: A Necessary and Sufficient Condition That Ordinary Least-Squares Estimations Be Best Linear Unbiased. J. American Statis. Assoc., Vol. 62, pp. 1302-1304, 1967.

4. Rao, C. R.: Least Squares Theory Using an Estimated Dispersion Matrix and Its Application to Measurement of Signals. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Univ. of Calif. Press (Berkeley, Calif.) Vol. 1, pp. 355-371, 1967.

5. Sebestyen, G. S.: Decision-Making Processes in Pattern Recognition. MacMillan Company (Yew York), 1962.

6. Anderson, T. W.: An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., 1958.

C:\Microfiche Scans\T1\1980020288\0001A01.tif,11/17/2009 2:29:22 PM,1,1,1,1,334,276,189,236
C:\Microfiche Scans\T1\1980020288\0001A02.tif,11/17/2009 2:29:25 PM,1,1,2,2,529,277,191,235
C:\Microfiche Scans\T1\1980020288\0001A03.tif,11/17/2009 2:29:29 PM,1,1,3,3,728,277,189,235
C:\Microfiche Scans\T1\1980020288\0001A04.tif,11/17/2009 2:29:32 PM,1,1,4,4,925,277,190,238
C:\Microfiche Scans\T1\1980020288\0001A05.tif,11/17/2009 2:29:33 PM,1,1,5,5,1121,277,190,236
C:\Microfiche Scans\T1\1980020288\0001A06.tif,11/17/2009 2:29:34 PM,1,1,6,6,1318,278,189,235
C:\Microfiche Scans\T1\1980020288\0001A07.tif,11/17/2009 2:29:34 PM,1,1,7,7,1514,278,190,236
C:\Microfiche Scans\T1\1980020288\0001A08.tif,11/17/2009 2:29:35 PM,1,1,8,8,1711,278,190,236
C:\Microfiche Scans\T1\1980020288\0001A09.tif,11/17/2009 2:29:36 PM,1,1,9,9,1908,278,190,236
C:\Microfiche Scans\T1\1980020288\0001A10.tif,11/17/2009 2:29:48 PM,1,1,10,10,2105,279,190,236