A THEORY OF LINEAR ESTIMATION

by

T. O. Lewis
Assistant Professor of Mathematics
Texas Technological College

P. L. Odell
Associate Director

Texas
Center for Research
3100 Perry Lane
Austin, Texas

The preparation of this monograph was supported in part by NASA Manned-Spacecraft Center, Contract No. 9-2619.

TABLE OF CONTENTS

Preface

The technique and theory for estimating unknown quantities or parameters from data are not new nor unknown to most mathematicians, biological scientists, physical scientists, and engineers. However, most developments of the theory usually assume that one knows or can approximate the probability density function of the random error which corrupts the data. Then the problem reduces to estimating the unknown parameters which define the density function.

Unfortunately, during several years as consultant industrial mathematicians we have found few who assume easily the form of the probability density function. The assumption of normal errors, that is, errors whose distribution is the normal probability density function, is indeed popular but sometimes dangerous.

We know of one case where the height of the hills in Korea were assumed normal. What this means is fun to conjecture. Even though the central limit theorem may support the assumption of normal errors there are certainly times when such an assumption is clearly false. Also, it is desirable to have a measure of reliability of ones estimate which (if one can get it) is based on an assumed probability density function. Usually we simply want to obtain the best estimate from the available data, hence no density function assumptions are necessary.

It is our intention here to develop a theory of linear estimation from a non-parametric (that is, with no assumptions concerning the underlying probability density functions associated with the errors in the data) point of view and indicate ways to extend this theory to problems in smoothing, filtering, extrapolation, and non-linear estimation.

Considerable attention, although not formally, is given to the concept of a robust estimator. A robust estimator is one that is good enough even though it is used in those instances where theoretically it does not apply.

The results here are those of a study which led to a rather large computer program for orbit determination. The technique used in the program was essentially the one of linear estimation. A large part of the material can be found in statistical literature, while the remainder is original with the authors and some colleagues at NASA - Manned Spacecraft Center, Houston, Texas.

We wish to acknowledge the contributions of Dr. H. P. Decell, Mr. Eugene Davis, Dr. Byron Tapley and A. H. Feiveson. We express our respect and appreciation to these colleagues who have taken time during the last three years to discuss these topics and their applications to various trajectory problems.

Chapter 1

MATHEMATICAL CONCEPTS

1.1 Matrices

The theory of linear statistical estimation that will be developed in this report will require some knowledge of the techniques of matrix algebra. In this chapter we will introduce those matrix concepts which perhaps may not be found in the usual undergraduate texts on matrix theory.

All matrices will be designated by capital English letters; the notation $A(m \times n) = (a_{ij})$ means that A is an $m \times n$ matrix having a_{ij} as the element in the i^{th} row and j^{th} column ($i = 1,\ldots,m$ = number of rows; $j = 1,\ldots,n$ = number of columns). All matrices considered will be presumed to have elements a_{ij} which are real numbers. A^* designates the transpose of A; thus if $A(m \times n) = (a_{ij})$ then $A^*(n \times m) = (b_{ij})$ where $b_{ij} = a_{ji}$. I_n is the $(n \times n)$ identity matrix with 1's down the diagonal and zeros elsewhere; usually the subscript n will be dropped if the dimension of I is clear from the context. E^n_{rs} is the $(n \times n)$ matrix (e_{ij}) wuch that $e_{rs} = 1$ and all other $e_{ij} = 0$; usually the superscript n will be dropped if the dimension of E_{rs} is clear from the context. A $(m \times n)$ is called square if $m = n$. If A is square then $|A|$ designates the determinant of A. If A is square and $|A| \neq 0$ then A^{-1} designates the inverse of A, and A is said to be nonsingular. ε_n designates the set of all real $(n \times 1)$ matrices; ε_n will also be called Euclidean n-space, and elements of ε_n called n dimensional vectors, or just vectors if the dimension is clear from the context. The symbol 0 will be used to designate either a matrix

which is identically zero or the scalar real number zero, depending upon the context. Given any x in ε_n, $||x|| \equiv \sqrt{x^*x}$ and is called the "norm' of x. The symbol "ϵ" will sometimes be used for "in" in the set theoretic sense; e.g. "given any $x\epsilon\varepsilon_n$" means given any x which is an element (a member) of the set ε_n. Let $W(m \times m)$ be positive definite so that $W = R^*R$ for some square R, $|R| \neq 0$, $|R| = R^*$. For any $a\epsilon\varepsilon_n$, define $||a||_W^2 \; a^*R^*Ra = ||Ra||^2$.

1.2 The Generalized Inverse of a Matrix

The importance of generalized inverses stems from the fact [4] that the matrix equation

$$SB = G$$

if consistent has a general solution given by

$$B = S^gB +)I - S^gS)Y,$$

where Y is an arbitrary matrix of appropriate dimensions. The matrix S^g is called a generalized inverse of S and has the property

$$SS^gS = S.$$

Four types of generalized inverses can be defined as follows:

DEFINITION 1.1 S^g is said to be a generalized inverse of S if

(1.1) $$SS^gS = S.$$

DEFINITION 1.2 S^r is said to be a reflexive generalized inverse of S if

(1.2) $$SS^rS = S \quad \text{and} \quad S^rSS^r = S^r.$$

DEFINITION 1.3 S^n is said to be a normalized generalized inverse of S if

(1.3) $$SS^nS = S, \quad S^nSS^n = S^n, \quad \text{and} \quad (SS^n)^* = SS^n.$$

DEFINITION 1.4 $S^\dagger$ is said to be the pseudoinverse of S if

(1.4) $$SS^\dagger S = S, \; S^\dagger SS^\dagger = S^\dagger, \; (SS^\dagger)^* = SS^\dagger, \text{ and } (S^\dagger S)^* = S^\dagger S.$$

The generalized inverse of Definition 1.1 has been studied by Bose [1] and Rao [6] with special reference to problems in least squares theory. The reflexive generalized inverse does not appear to have been studied although its existence was pointed out by Rao, and Frame [2] has indicated an equivalent type under the term semi-inverse. The normalized generalized inverse was introduced by Zelen and Goldman [8] who used the term weak generalized inverse. The pseudo-inverse, which is unique, was first introduced by Moore [3] and later by Penrose under the names general reciprocal and generalized inverse, respectively. Computational aspects have been studied by various authors [2], [6], [7], [8].

1.3 Some Properties of Generalized Inverses

If we let G_g, G_r, G_n, and $G_\dagger$ denote the set of all generalized, reflexive generalized, normalized generalized and pseudoinverses of a

matrix, then it is clear from the definitions that

$$G_{\dagger} \subset G_n \subset G_r \subset G_g$$

The present investigation of further relationships between, and properties of, the elements of these sets is based on examining the relationship of a property of a matrix S to the corresponding property of a typical element of G_i, $i = g, r, n, \dagger$. The particular characteristics selected in this investigation are rank, symmetry, characteristic roots, and characteristic vectors.

THEOREM 1.1 If S^g is any generalized inverse of S then

$$\text{rank } (S^g) \geq \text{rank } (S) = \text{rank } (S^gS) = \text{rank } (SS^g).$$

Proof. Since the rank of a product does not exceed the rank of either factor the conclusions follow from the equations

$$\text{rank } (S) \geq \text{rank } (SS^g) \geq \text{rank } (SS^gS) = \text{rank } (S),$$

$$\text{rank } (S) \geq \text{rank } (S^gS) \geq \text{rank } (SS^gS) = \text{rank } (S),$$

and

$$\text{rank } (S^g) \geq \text{rank } (SS^gS) = \text{rank } (S).$$

In many investigations, it is desired that the rank of S^g be the same as the rank of S. The following theorem indicates a necessary and sufficient condition for this to hold.

THEOREM 1.2 A necessary and sufficient condition that rank (S) = rank (S^g) is that S^g be a reflexive generalized inverse of S.

Proof. If S^g is a reflexive generalized inverse of S an application of Theorem 1.1 to S^g shows that rank (S^g) $\leq$ rank (S) and hence (S^g) = rank (S).

Let S be an n X p matrix of rank r. If rank (S) = rank (S^g), simple matrix multiplication shows that if S is expressed as $S = P_1^{-1}BP_2^{-1}$ (which is always possible), where

$$B = \left[\begin{array}{c|c} I(r) & 0_{r,p-r} \\ \hline 0_{n-r,r} & 0_{n-r,p-r} \end{array}\right] \tag{1.5}$$

and where I(r) is the r X r identity matrix and $0_{r,p-r}$ is the r X (p-r) null matrix, then a generalized inverse must be of the form

$$P_2B^rP_1, \text{ where } B^r = \left[\begin{array}{c|c} I(r) & V \\ \hline W & WV \end{array}\right] .$$

Since $P_2B^rP_1$ can be shown to be a reflexive generalized inverse of S the conclusion follows.

It follows from Theorem 1.2 and the definitions that normalized generalized inverses S^n and the pseudoinverse $S^\dagger$ have the same rank as S. In addition the representation $S = P_1^{-1}BP_2^{-1}$ with B given by (1.5) shows that rank (S^g) can assume any of the values rank (S),...,n by appropriately choosing a generalized inverse of B.

If S is Hermitian, then it seems reasonable to inquire into the Hermitian nature (if any) of the various types of generalized inverses

of S. Examples of generalized inverses of 2 X 2 Hermitian matrices easily show that generalized inverses, reflexive generalized inverses, and normalized generalized inverses need not be Hermitian. For example, if

$$S = \begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix},$$

then a generalized inverse, reflexive generalized inverse, and a normalized generalized inverse are given by

$$S^g = \begin{bmatrix} 1-2i & -1 \\ 1 & 0 \end{bmatrix}, \quad S^r = \begin{bmatrix} 1+i & 1 \\ 0 & 0 \end{bmatrix}, \quad S^n = \begin{bmatrix} 1/2 & i/2 \\ 0 & 0 \end{bmatrix}.$$

The expression $1/2\ (S^g + S^{g*})$ shows that a Hermitian generalized inverse always exists if S is Hermitian. Similarly

$$S^r = \begin{bmatrix} S_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

where

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^* & S_{22} \end{bmatrix}$$

and $(S_{11} \quad S_{12})$ is a basis for the rowspace of S, shows that a Hermitian reflexive generalized inverse always exists if S is Hermitian. Note that

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^* & S_{12}^* S_{11}^{-1} S_{12} \end{bmatrix}.$$

It is easy to show that the pseudoinverse of a Hermitian matrix is Hermitian (in fact, Penrose [4] proved the stronger result that S normal implies $S^\dagger$ normal). If S is Hermitian and S^n is Hermitian then it can be shown that $S^n \equiv S^\dagger$.

The characteristic roots and vectors of a matrix are of interest in many investigations of matrices. It is well-known that the characteristic vectors of S and S^{-1} are identical, with the corresponding characteristic roots reciprocal. It is of interest, therefore, to determine to what extent (if any) generalized inverses display this same behavior. It is easy to construct examples of generalized inverses S^i, $i = g, r, n, \dagger$, which have different characteristic vectors than S and also examples for which the characteristic roots are not reciprocals, so that additional assumptions on S are necessary.

We define properties R and V as follows:

A generalized inverse will be said to have property R if the reciprocals of nonzero characteristic roots of S are characteristic roots of S^g and conversely.

A generalized inverse will be said to have property V if x is a characteristic vector of S with root λ implies that x is a characteristic vector of S^g with λ^{-1} and conversely.

Obviously property R is weaker than property V. Remarks made previously indicate that generalized inverses of all types do not necessarily possess properties R or V.

THEOREM 1.3 If S is Hermitian then S^n possesses property R.

Proof. Let λ be a nonzero characteristic root of S and x its associated characteristic vector. Multiplying both sides of the

equation $Sx = \lambda x$ by SS^n yields

$$x = SS^n x.$$

Since $SS^n = S^{n*}S$ we have

$$\bar{\lambda}^{-1}x = S^{n*}x.$$

Hence $\bar{\lambda}^{-1}$ is a characteristic root of S^{n*} and hence $\bar{\lambda}^{-1}$ is a characteristic root of S^n.

Conversely let η be a nonzero characteristic root of S^n and y be the associated characteristic vector. Multiplying both sides of the equation $S^n y = \eta y$ by S and using the fact that $(SS^n) = S^{n*}S$ shows that

$$S^{n*}(Sy) = \eta(Sy).$$

Hence

$$\eta^{-1}(Sy) = S(Sy).$$

Thus Sy is a characteristic vector of S with characteristic root η^{-1}. Note that if $Sy = 0$ then $S^{n*}Sy = SS^n y = 0$ or $S^n y = 0$, which is a contradiction.

The following lemma gives a sufficient condition for a reflexive generalized inverse to possess property V.

LEMMA 1.1 If S and S^r commute then S^r possesses property V.

Proof. If S commutes with S^r then the conclusion follows from

$$Sx = \lambda x \Rightarrow Sx = \lambda SS^r x \Rightarrow \lambda^{-1}x = S^r x,$$

$$S^r y = \eta y \Rightarrow S^r y = \eta S^r Sy \Rightarrow \eta^{-1}y = Sy.$$

This lemma is related to a result of Price [5] who proved that if S^n and $S^\dagger$ commute for some 1 then if x is a characteristic vector of S with nonzero characteristic root λ it follows that x is also a characteristic vector of $S^\dagger$ with characteristic root λ^{-1}.

THEOREM 1.4 If S is normal, then $S^\dagger$ possesses property V. Conversely, if S and S^g are normal and S^g possesses property V, then $S^g = S^\dagger$.

Proof. By Penrose's Lemma 1.8 we have $SS^\dagger = S^\dagger S$ and the conclusion follows from Lemma 1.1.

Normality of S, S^g and property V imply that the spectral representations of S and S^g are $\sum_i \lambda_i E_i$ and $\sum_i \lambda_i^{-1} E_i$, respectively, and one easily verifies that $S^g = S^\dagger$.

We note that Theorem 1.2 provides a partial characterization of a reflexive generalized inverse in terms of rank and that Theorem 1.4 provides a similar characterization (under the assumption that S is normal) of the pseudoinverse in terms of characteristic vectors. No such characterization has yet been obtained for the normalized generalized inverses although Zelen and Goldman [8] have established that S^n is a normalized generalized inverse if and only if it can be written in the form $S^n = (S^*S)^g S^*$ where $(S^*S)^g$ is a generalized inverse of S^*S.

Some additional well-known properties of the pseudoinverse are given below:

THEOREM 1.5 For every (m X n) matrix A, there exists a unique (n X m) matrix, which we shall designate as A^+, that satisfied the following four identities:

(1) $AA^+A = A$

(2) $A^+AA^+ = A^+$

(3) $(AA^+)^* = AA$ (m X m)

(4) $(A^+A)^* = A^+A$(n X n)

Furthermore,

(R1) if $D = (d_{ij})$ is square (m = n) and diagonal ($d_{ij} = 0$ for $i \neq j$) then $D^+ = (d^+_{ij})$ is defined by $d^+_{ij} = 0$ for $i \neq j$, $d^+_{ii} = 0$ if $d_{ii} = 0$, $d^+_{ii} = d^{-1}_{ii}$ if $d_{ii} \neq 0$.

(R2) if $A^*A = PDP^*$, where $PP^* = P^*P = I$, and D is diagonal, then $A^+ = PD^+P^*A^*$.

(R3) if A = BC, where the columns of B are linearly independent and the rows of C are linearly independent, then $A^+ = C^*(CC^*)^{-1}(B^*B)^{-1}B^*$. Thus

(R3.1) $A^+ = (A^*A)^{-1}A^*$ if the columns of A are linearly independent.

(R3.2) $A^+ = A^*(AA^*)^{-1}$ if the rows of A are linearly independent.

(R3.3) $A^+ = A^{-1}$ if A is square and nonsingular.

THEOREM 1.6 The matrix correspondence A -- A^+ satisfies the following:

(R1) $(A^+)^+ = A$

(R2) $(A^*)^+ = (A^+)^* \equiv A^{+*} \equiv A^{*+}$

(R3) $A^+AA^* = A^*$

(R4) $A^*AA^+ = A^*$

(R5) $AA^{+}A^{+*} = A^{+*}$

(R6) $A^{+*}A^{+}A = A^{+*}$

(R7) $A^{*+}A^{*}A = A$

(R8) $AA^{*}A^{*+} = A$

(R9) $A^{*}A^{+*}A^{+} = A^{+}$

(R10) $A^{+}A^{+*}A^{*} = A^{+}$

(R11) The row spaces of A^{+} and A^{*} are identical, i.e., the rows of A^{+} are in the row space of A^{*} and the rows of A^{*} are in the row space of A^{+}.

(R12) The column spaces of A^{+} and A^{*} are identical.

(R13) A, A^{+} and A^{*} all have the same rank.

(R14) $(AA^{*})^{+} = A^{+*}A^{+}$.

(R15) $(AA^{*})^{+}(AA^{*}) = AA^{+}$.

(R16) If A^{+} commutes with some power of A and λ is any non-zero eigen value of A corresponding to the eigen vector x, then λ^{-1} is an eigen value of A^{+} corresponding to the eigen vector x.

(R17) If $\alpha \neq 0$ then $(\alpha A)^{+} = \alpha^{-1}A^{+}$.

(R18) $0^{+} = 0^{*}$.

DEFINITION for $A(m \times n)$ and $b(m \times 1)$,

$$(A,b) \equiv A^{+}b + (I_n - A^{+}A)\xi_n$$

LEMMA 1.2 The element of least norm in (A,b) is $A^{+}b$.

LEMMA 1.3 Let $x \in \xi_n$. Then $x \in (A,b)$ if and only if $A(x - A^{+}b) = 0$.

LEMMA 1.4 Let $x \in \xi_n$. Then $x \in)A,b)$ if and only if $A^{*}(Ax - b) = 0$.

LEMMA 1.5 Let $A(m \times n)$, $N(n \times n)$ nonsingular, $M(m \times m)$ nonsingular. Then $(AN)(AN)^+ = AA^+$ and $(MA)^+(MA) = A^+A$.

LEMMA 1.6 Let $A(m \times n)$, and $N(n \times n)$ nonsingular. Then $(A,b) = N\ (AN,b)$.

LEMMA 1.7 Let $V = S^2$ be positive definite $(n \times n)$, $A(m \times n)$, $b(m \times 1)$. Then the vector $(n \times 1)x$ of least $||x||_V$ in (A,b) is given by $S^{-1}(AS^{-1})^+b$.

LEMMA 1.8 The following statements are equivalent:

(1) The columns of A are linearly independent.

(2) A^*A is nonsingular.

(3) $A^+A = I$.

THEOREM 1.7 The equation $Ax = b$ has a solution (vector) x if and only if $AA^+b = b$. If the latter equality holds then x is a solution if and only if $x \in (A,b)$.

THEOREM 1.8 (Least Squares). For $A(m \times n)$ and $b(m \times 1)$, the set of all $(n \times 1)$ vectors x such that $||Ax - b||$ is a minimum, is (A,b). Also, the $n \times 1$ matrix (vector) of least norm such that $||b - Ax||$ is minimized, is A^+b.

COROLLARY 1.8.1 For $A(m \times n)$ and $b(m \times 1)$, and $W(m \times m) = R^*R$ which is positive definite, the set of all $(n \times 1)$ vectors such that $||Ax - b||_W$ is a minimum, is (RA,Rb). The vector of least norm such that $||Ax - b||_W$ is minimized, is $(RA)^+Rb$.

COROLLARY 1.8.2 Let $V = S^2$ be positive definite $(n \times n)$, $W = R^2$ positive definite $(m \times m)$, $A(m \times n)$, $b(m \times 1)$. Then the set of all $(n \times 1)$ vectors such that $||Ax - b||_W$ is a minimum, is (RA,Rb).

The vector of least "V" norm such that $||Ax - b||_W$ is minimized, is $S^{-1}(RAS^{-1})^{+}Rb$.

THEOREM 1.9 Let A be m X n and Z be any m X 1 matrix (vector). Then there exist m X 1 matrices (vectors) x and y such that

(1) $z = x + y$

(2) x is in the column space of A

(3) y is orthogonal to the column space of A

Any vectors satisfying (1)-(3) above are unique, and

(4) $x = AA^{+}z$

(5) $y = z - AA^{+}z$

(6) $x^{*}y = 0$

Thus AA^{+} is the projection which takes any column vector (m X 1) into the column space of A; $I_m - AA^{+}$ is the projection which takes any (m) vector into the orthogonal complement of the column space of A.

THEOREM 1.10 For the matrix equation A X B = C to have a solution, a necessary and sufficient condition is

$$AA^{+}CB^{+}B = C$$

in which case, the general solution is

$$X = A^{+}CB^{+} + Y - A^{+}AYBB^{+}$$

where Y is arbitrary to within the limits of being consistent with the demension in the indicated multiplications.

Proof: If X satisfied A X B = C,

$$C = A\,X\,B = AA^{+}A\,X\,BB^{+}B = AA^{+}CB^{+}B.$$

Conversely, if $C = AA^+CB^+B$, A^+CB^+ is a particular solution. Clearly, for the general solution $A\,X\,B = \emptyset$ must be solved. Any expression of the form

$$X = Y - A^+A\,Y\,BB^+$$

is a solution. The only property required property of A^+ and B^+ is

$$AA^+A = A$$

$$BB^+B = B$$

COROLLARY 1.6.1 A necessary and sufficient condition for the equations

$$Ax = c$$

is

$$x = A^+c + (I - A^+A)y$$

where y is arbitrary, provided a solution exists.

THEOREM 1.11 A^+A, AA^+, $I - A^+A$ and $I - AA^+$ are hermitian idempotent. If H is hermitian idempotent, then $H^+ = H$.

Proof: The proof requires a straightforward application of Theorem 1.1.

In general, the reversal rule, $(AB)^+ = B^+A^+$ as in the case of the standard inverse, does not hold. R. Cline [11] obtained the following results.

THEOREM 1.12 Let A and B be matrices with the produce AB defined. Then,

$$(AB)^+ = B_1{}^+A_1{}^+$$

where:

$$AB = A_1B_1$$

$$B_1 = A^+AB$$

$$A_1 = AB_1B_1^+$$

Proof: The produce AB can be written as

$$AB = AA^+AB = AB_1 = AB_1B_1^+B_1 = A_1B_1.$$

Let $y = AB = A_1B_1$ and let $x = B_1^+A_1^+$. Then it is only necessary to show that y and x satisfy the equations in Definition 1.4. From the definition of A, we have that $A_1B_1B_1^+ = AB_1B_1^+B_1B_1^+ = A_1$. Now $yx = A_1B_1B_1^+A_1^+ = A_1A_1^+$ is hermitian. Also $yxy = A_1B_1B_1^+A_1^+A_1B_1 = A_1A_1^+A_1B_1 = A_1B_1 = y$ and $xyx = B_1^+A_1^+(A_1B_1B_1^+)A_1^+ = B_1^+A_1^+A_1A_1^+ = B_1^+A_1^+ = x$. In order to show that xy is hermitian, we observe first that using the definitions of A_1 and B_1 that

$$A^+A_1 = A^+AB_1B_1^+ = A^+A(A^+AB)B_1^+ = A^+ABB_1^+ = B_1B_1^+.$$

Also, since $A_1^+A_1A_1B_1^+ = A_1A_1^+$, with both A_1^+A and $B_1B_1^+$ hermitian, $B_1B_1^+A_1^+A_1 = A_1^+A_1$. Substituting A^+A_1 for $B_1B_1^+$ gives $A_1^+A_1 = A^+A_1A_1^+A_1 = A^+A_1$ and so $A_1^+A_1 = B_1B_1^+$.

From this it now follows that $xy = B_1^+A_1^+A_1B = B_1^+B_1B_1^+B_1 = B_1^+B_1$ is hermitian. Since it has been shown that y and x satisfy the defining equations for the pseudoinverse, $x = y^+$. But $x = B_1^+A_1^+$.

It is of interest to show that $(AB)^{+} = B^{+}A^{+}$ under certain conditions imposed upon A and B. The following theorems were obtained by T. M. E. Greville [10]. In fact, he defines $C^{\dagger}$ as the unique matrix satisfying these two equations.

THEOREM 1.13 If A and B are otherwise arbitrary matrices such that AB is defined, $(AB)^{+} = B^{+}A^{+}$ if and only if both the equations

(1.6) $$A^{+}ABB^{*}A = BB^{*}A^{*}$$

and

(1.7) $$BB^{+}A^{*}AB = A^{*}AB$$

are satisfied.

Proof: Multiplying (1.6) on the left by B^{+} and on the right by $(AB)^{*+}$, using the fact that $C^{+}CC^{*} = C^{*}C^{+}C = C^{*}$ and using the fact that $CC^{*}C^{*+} = C^{*+}C^{*}C = C$ in the form

$$(AB)(AB)^{*}(AB)^{*+} = AB$$

gives

(1.8) $$B^{+}A^{+}AB = (AB)^{*}(AB)^{*+} = (AB)^{+}(AB).$$

Similarly, taking transposes of both sides of (1.7) gives

(1.9) $$B^{*}A^{*}ABB = B^{*}A^{*}A,$$

and then multiplying on the right by A^{+} and on the left by $(AB)^{*+}$ and using the fact that $C^{+}CC^{*} = C^{*}CC^{+} = C^{*}$ and $CC^{*}C^{*+} = C^{*+}C^{*}C = C$ leads to the equation

(1.10) $\quad ABB^{+}A^{+} = AB(AB)^{+}.$

In view of the fact that CC^{+} and $C^{+}C$ are projection operators on $R(C)$ and $R(C^{*})$ respectively, we find that (1.8) and (1.10) express the fact that $B^{+}A^{+}$ is the generalized inverse of AB, as defined by Moore.

Conversely, $(AB)^{+} = B^{+}A^{+}$ implies

$$B^{*}A^{*} = B^{+}A^{+}ABB^{*}A^{*}.$$

Multiplying on the left by $ABB^{*}B$ and using $B^{*}BB^{+} = B^{*}$ gives

$$ABB^{*}(I - A^{+}A)BB^{*}A^{*} = \theta,$$

where θ denotes a null matrix. As the left member is Hermitian and $I - A^{+}A$ is idempotent, it follows that

$$(I - A^{+}A)BB^{*}A^{*} = \theta,$$

which is equivalent to (1.6). In an analogous manner (1.7) is obtained.

THEOREM 1.14 $(AB)^{+} = B^{+}A^{+}$ if and only if both $A^{+}ABB^{*}$ and $A^{*}ABB^{+}$ are Hermitian.

Proof: If $A^{+}ABB^{*}$ is Hermitian, we have

$$A^{+}ABB^{*} = BB^{*}A^{+}A,$$

and multiplication on the right by A^{*} gives (1.6). Conversely, multiplication of (1.6) on the right by A^{*+} gives

(1.11) $\quad A^{+}ABB^{*}A^{+}A = BB^{*}A^{+}A.$

Since the left member of (1.11) is Hermitian, the right member is also.

In a similar fashion it can be shown that (1.7) is equivalent to the statement that A^*ABB^+ is Hermitian.

It will be noted that an equivalent statement to the condition in Theorem 1.14 is that A^+A and BB^* commute and also A^*A and BB^+ commute.

THEOREM 1.15 $(AB)^+ = B^+A^+$ if and only if

$$A^+ABB^*A^*ABB^+ = BB^*A^*A. \tag{1.12}$$

Proof: Multiplying (1.12) on the left by A^+A gives

$$A^+ABB^*A^*ABB^+ = A^+ABB^*A^*A. \tag{1.13}$$

Combining (1.12) and (1.13) gives

$$A^+ABB^*A^*A = BB^*A^*A,$$

and multiplication on the right by A^+ gives (1.6). An analogous process leads to (1.9), which is equivalent to (1.7).

On the other hand, if (1.6) and (1.7) hold, multiplying (1.6) on the right by A and then using (1.9) to transform the left member gives (1.12).

Equations (1.6) and (1.7) have a simple interpretation in terms of range spaces. They assert, respectively, that (A^*) is an invariant space of BB^* and that (B) is an invariant space of A^*A. In some particular cases this interpretation leads to a characterization of those matrices B that satisfy $(AB)^+ = B^+A^+$ for a given A.

For example, if A is of full column rank, $A^+A = I$ and (1.6) is immediately satisfied. Then (1.7) holds if and only if B is a null matrix or (B) is the space spanned by some set of eigen vectors of A^*A.

THEOREM 1.16 $(AB)^+ = B^+A^+$ if and only if both the equations

(1.14) $$A^+AB = B(AB)^+AB$$

and

(1.15) $$BB^+A = A^*AB(AB)^+$$

are satisfied.

Proof: Multiplication of (1.6) on the right by $(AB)^{*+}$ gives (1.14), and conversely multiplication of (1.14) on the right by $(AB)^*$ gives (1.6). Similarly it can be shown that (1.15) is equivalent to (1.9).

THEOREM 1.17 A necessary condition for $(AB)^+ = B^+A^+$ is that A^+A and BB^+ commute.

Proof: Substitution of B^+A^+ for $(AB)^+$ in (1.14) and multiplication on the right by B^+ gives

$$A^+ABB^+ = BB^+A^+ABB^+.$$

As the right member is Hermitian, the conclusion follows.

That the condition of Theorem 1.17 is not sufficient is clear from the example:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad (AB)^+ = (0 \quad 1), \quad B^+A^+ = (-1 \quad 1).$$

As A is nonsingular, $A^+A = A^{-1}A = I$, and the condition is fulfilled.

It is easily seen that the commutativity of A^+A and BB^+ is equivalent to either of the conditions

$$A^+ABB^+A^* = BB^+A^*$$

and

$$BB^+A^+AB = A^+AB.$$

These equations can be interpreted as asserting that (A^*) is the direct sum of a subspace of (B) and a space orthogonal to (B) and that (B) is the direct sum of a subspace of (A^*) and a space orthogonal to (A^*). These observations reveal something about the structure of matrices A and B that satisfy $(AB)^+ = B^+A^+$. It is easily seen that (1.6) and (1.7) are equivalent to the following two equations:

$$\begin{aligned} &(I - A^+A)BB^*A^+A = \theta. \\ &(I - BB^+)A^+ABB^+ = \theta. \end{aligned} \tag{1.16}$$

Equation (1.16) shows that if B is resolved into the two component matrices,

$$B_1 = A^+AB, \quad B_2 = (I - A^+A)B,$$

then not only do we have $B_1^*B_2 = \theta$ as expected, but also $B_2B_1^* = \theta$. Similar remarks apply to the resolution of A^* into

$$A_1^* = BB^+A^*, \quad A_2^* = (I - BB^+)A^*.$$

1.4 Quadratic Forms

DEFINITION 1.5 If X is an $n \times 1$ vector whose elements are in the complex field, then the complex type quadratic form X^*AX is defined as $\sum_{i=1}^{n} \sum_{j=1}^{n} \bar{x}_i x_j a_{ij}$. Similarly if the elements of X are in real field, then the real type quadratic form X^TAX is defined as $\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j a_{ij}$.

The following definition will be given only for the complex type quadratic forms. Similar definitions hold for the real type quadratic forms.

DEFINITION 1.6 The rank of the quadratic form X^*AX is the rank of the matrix A.

DEFINITION 1.7 The quadratic form X^*AX is said to be positive definite if and only if $X^*AX > 0$ for all vectors $X \neq 0$.

DEFINITION 1.8 The quadratic form X^*AX is said to be positive semi-definite if and only if $X^*AX \geq 0$ for all vectors X.

THEOREM 1.18 A necessary and sufficient condition for a Hermitian (symmetric) matrix A to be positive definite is that there exists a nonsingular matrix P such that $A = P^*P(P^TP)$.

DEFINITION 1.9 A characteristic root of a $n \times n$ matrix A is a scalar λ such that $AX = \lambda X$ for some vector $X \neq \emptyset$. The vector X is called the characteristic vector of the matrix A.

A necessary and sufficient condition for an eigen vector to exist is that there should be a solution of $(A - \lambda I)X = 0$ for which $X \neq 0$.

Such a solution will exist if and only if $\det(A - \lambda I) = 0$. Since the $\det(A - \lambda I)$ is a polynomial of the n^{th} degree in λ, it will certainly have a zero, real or complex.

THEOREM 1.19 The number of nonzero characteristic values of a matrix A is equal to the rank of A.

THEOREM 1.20 The characteristic roots of a Hermitian (symmetric) matrix are real.

THEOREM 1.21 The characteristic values of a positive definite matrix A are positive; the characteristic values of a positive semidefinite matrix are non-negative.

THEOREM 1.22 For every symmetric matrix A there exists an orthogonal matrix P such that $P^TAP = D$, where D is a diagonal matrix whose diagonal elements are the characteristic roots of A.

It is sometimes advantageous to break a matrix into submatrices. This is called partitioning a matrix into submatrices. The following example will illustrate the above. Let A be an $n \times n$ matrix and write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} is $n_1 \times m_1$, A_{12} is $n_1 \times (n - m_1)$, A_{21} is $(n - n_1) \times m_1$, and A_{22} is $(n - n_1) \times (n - m_1)$.

The product AB of two matrices can be made symbolically even if A and B are broken into submatrices. The multiplication proceeds as

if the submatrices were single elements of the matrix. However, the dimensions of the matrices and of the submatrices must be such that they will multiply.

THEOREM 1.23 If A is a positive definite symmetric matrix such that

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

and if B is the inverse of A such that

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

and if B_{ii} and A_{ii} are each of dimension m_i X m_i, etc. then

$$\begin{aligned} B_{11}^{-1} &= A_{11} - A_{12}A_{22}^{-1}A_{21} \\ B_{22}^{-1} &= A_{22} - A_{21}A_{11}^{-1}A_{12} \\ B_{12} &= - A_{11}^{-1}A_{12}B_{22}^{-1} \\ B_{21} &= - A_{22}^{-1}A_{21}B_{11}^{-1} \end{aligned} \tag{1.17}$$

THEOREM 1.24 If $P_1 = P_0 + A^*A$ where P_0 is an (n X n) positive definite matrix and A is any r X m matrix, then

$$P_1^{-1} = P_0^{-1} - P_0^{-1}A^*(AP_0^{-1}A^* + I)^{-1}AP_0 \tag{1.18}$$

Proof: Since P_0 is positive definite, then P_0^{-1} is positive definite. Hence it follows $(AP_0^{-1}A^* + I)$ is positive definite which

implies $(AP_0^{-1}A^* + I)^{-1}$ exists. Therefore

$$P_1P_1^{-1} = I + P_0^{-1}A^*A - P_0^{-1}A^*(AP_0^{-1}A^* + I)^{-1}(AP_0^{-1}A^* + I)^{-1}A$$

$$= I.$$

The above inversion formula (1.18) has been used extensively in the sequential estimation theory for updating estimates as more samples are observed [12] and [13]. It is of interest to know when a formula similar to (1.18) holds for pseudoinverses since in some applications P_0 may be Hermitian positive semi-definite.

THEOREM 1.25 If P_0 is a positive semi-definite (Hermitian) $m \times m$ matrix and A is an $r \times m$ matrix with $P_1 = P_0 + A^*A$ then

$$(1.19) \qquad P_1^+ = P_0^+ - P_0^+A^*(AP_0^+A^* + I)^{-1}AP_0^+$$

if, and only if, the null space of A contains the null space of P_0.

Proof: Since P_0 is positive semi-definite, then P_0^+ is positive semi-definite which implies $x^*(AP_0^+A^* + I)x \geq 0$, $x \in X_m$ and $x^*(AP_0^+A^* + I)x = 0$ if and only if $x = \phi$. Hence $(AP_0^+A^* + I)^{-1}$ exists. Suppose $N(A) \supset N(P_0)$. Then $N(P_1) = N(P_0)$ since $N(P_1) = N(P_0) \cap N(A)$. To show that $P_1^+P_1x = x$, for each $x \in R(P_0^+)$. Let $x \in R(P_0^+)$, then $P_1^+P_1x = x + P_0^+A^*Ax - P_0^+A^*(AP_0^+A^* + I)^{-1}(AP_0^+A^* + I)$. $Ax = x$. Since $X_m = N(P_0) + R(P_0^+)$, then $x \in X_m$ can be written as $x = x_1 + x_2$ where $x_1 \in R(P_0^+)$ and $x_2 \in N(P_0)$. Thus $P_1^+P_1x = P_1^+P_1x_1 = x$. Hence $P_1^+P_1$ is a projection operator on $R(P_0^+)$.

Conversely suppose $P_1^+ = P_0^+ - P_0^+A^*(AP_0^+A^* + I)^{-1}AP_0^+$. Thus it follows $R(P_1^+) \subset R(P_0^+)$ and $N(P_1) = N(P_0) \cap N(A)$ which implies

$R(P_1^+) \subset R(P_0^+)$ and $N(P_1) \subset N(P_0)$. But the only way this can be true is for $R(P_1^+) = R(P_0^+)$ and $N(P_1) = N(P_0)$ which implies $N(A) \supset N(P_0)$.

To show that $(P_1P_1^+)^* = P_1P_1^+$ and $(P_1^+P_1)^* = (P_1^+P_1)$ observe that

$$\begin{aligned}(P_1P_1^+)^* &= [P_0P_0^+ + A^*AP_0^+ - A^*(AP_0^+A^* + I)(AP_0^+A^* + I)^{-1}AP_0^+]^* \\ &= P_0P_0^+ \\ &= P_1P_1^+.\end{aligned}$$

A similar argument shows $(P_1^+P_1)^* = P_1^+P_1$.

1.5 The Crout Factorization

Let $P = (p_{ij})$, $i,j=1,2,\ldots,n$, be positive definite, real, symmetric matrix. It is shown in Gantmacher [15], that P has a factorization

$$P = TT^T, \tag{1.20}$$

where T is lower-triangle, with positive elements on the main diagonal. If the existance of the factorization (1.20) is given, then it is easy to show how to compute the components t_{ij} of T in the order

$$ij = 11, 21, \ldots, n1; 22, 32, \ldots, n2; \ldots; nn.$$

Since $t_{ij} = 0$ for $j > i$ (1.38) states that

$$p_{ij} = \sum_{k=1}^{j} t_{ik}t_{jk}. \tag{1.21}$$

First we compute

$$t_{11} = p_{11}^{1/2}. \tag{1.22}$$

The other elements in the first column are

$$t_{i1} = t_{11}^{-1} p_{i1}, \quad i = 2,3,\ldots,n. \tag{1.23}$$

If the preceding columns $k < j$ have been computed, we compute the diagonal element.

$$t_{jj} = (p_{jj} - \sum_{k=1}^{j-1} t_{jk}^2)^{1/2}. \tag{1.24}$$

If $j < n$, the elements below the diagonal are computed from the formula

$$t_{ij} = t_{jj}^{-1}(p_{jj} - \sum_{k=1}^{j-1} t_{ik}t_{jk}), \; i = j + 1, \ldots, n. \tag{1.25}$$

1.6 References

[1] R. C. Bose, Lecture Notes on Analysis of Variance, University of North Carolina, Chapel Hill, North Carolina, 1959.

[2] J. S. Frame, Matrix Functions and Applications, IEEE Spectrum, 1 (March 1964), pp. 208-220.

[3] E. H. Moore, Abstract, Bull, Amer. Math. Soc., 26 (1920), pp. 394-395.

[4] R. Pinerose, A Generalized Inverse for Matrices, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406-418.

[5] C. M. Price, The Matrix Pseudoinverse and Minimal Variance Estimates,

[6] C. R. Rao, A Note on a Generalized Inverse of a Matrix with Applications to Problems in Mathematical Statistics, J. Royal Statist. Soc. Ser. B, 24 (1962), pp. 152-158.

[7] C. A. Rohde, Contributions to the Theory, Computation and Applications of Generalized Inverses, Ph.D. Dissertation, North Carolina State University, Raleigh, 1964.

[8] M. Zelen and A. J. Goldman, Weak Generalized Inverses and Minimum Variance Linear Unbiased Estimation, Tech. Report 314, Mathematics Research Center, United Stated Army, University of Wisconsin, Madison, 1963.

[9] Charles A. Rohde, Some Results on Generalized Inverses, Siam Review, Vol.8, No.2, April 1966.

[10] T.N.E. Greville, The Pseudo-Inverse of a Rectangular Matrix and its Applications to the Solution of Systems of Linear Equations, Siam Review, 1 (January 1959), pp. 38-43.

[11] R. E. Cline, Note on the Generalized Inverse of the Product of Matrices, Siam Review, 6 (January 1964), pp. 57-58.

[12] Yu Chi Ho, On the Stochastic Approximation Method and Optimal Filtering Theory, Jour. Math. Analysis and App. 6, (1962), pp. 152-154.

[13] Ralph Deutsch, Estimation Theory, Prentice Hall, Englewood Cliffs, New Jersey (1965).

[14] L. A. Zadeh and C. A. Desorer, Linear System Theory, McGraw-Hill, New York (1963).

[15] F. R. Gantmacher, The Theory Matrices, Vols 1 and 2, transl., K. A. Hirsch, Chelsea, New York, 1959.

[16] Joel N. Franklin, Numerical Simulation of Stationary and Non Stationary Gaussian Random Processes, Siam Review, Vol. 7, No. 1, January 1965.

Chapter II

MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

In this chapter we will formulate and prove the well-known [1], [2], [3] Gauss Markov Theorem. Briefly, the theorem directs our attention to a simple form of a minimum variance linear estimator which is remarkably applicable to most any kind of estimation problem. It has at least two forms and is used directly and indirectly in almost every field of the sciences in which data is collected to estimate a parameter.

2.1 The Classical Form of the Gauss Markov Theorem

The theorem as is usually stated is as follows:

THEOREM 2.1 Let $y = Hx+v$ be a linear statistical model, where y is a $p \times 1$ vector of observations; H is a $p \times n$ known mapping matrix of rank $p \leq n$; x is a $n \times 1$ unknown state (parameter) vector and v is a $p \times 1$ random vector such that

$$Ev = \phi$$

$$E\sigma v^T = R,$$

a positive definite covariance matrix. Then the minimum variance linear unbiased estimator of x, denoted by $\hat{x}$ is given by

$$\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} y. \tag{2.1}$$

Proof: Since we require that $\hat{x}$ to be linear and unbiased then $\hat{x}$ must be of the form $\hat{x} = By$ and $E\hat{x} = x$, respectively on selecting

$$B = (H^T R^{-1} H)^{-1} H^T R^{-1} \tag{2.2}$$

we see that $\hat{x}$ is indeed linear.

Consider $E(\hat{x}) = E[(H^T R^{-1} H)^{-1} H^T R^{-1} y] = (H^T R^{-1} H)^{-1} H^T R^{-1} E\{y\}$

$$= (H^T R^{-1} H)^{-1} H^T R^{-1} E(Hx+v) = (H^T R^{-1} H)^{-1} H^T R^{-1} Hx = x.$$

Hence, $\hat{x}$ is unbiased.

Let x^* be any linear estimator of x. We can write x as

$$x^* = B^* y$$

where B^* is a mapping matrix. Without loss of generality we can write

$$B^* = B + C,$$

where B is defined by (2.2) and C is the residual matrix $B^* - B$. Then

$$x^* = \hat{x} + CY. \tag{2.3}$$

We require that x^* to be unbiased, that is,

$$x = E\, x^* = E(\hat{x} + CY)$$
$$= E(\hat{x} + C[Hx + v)$$
$$= x + CHx,$$

which in turn requires that

$$CH = \phi. \tag{2.4}$$

Consider the covariance matrix of x^* denoted here by

$$C(x^*,x^{*T}) = E(x^* - x)(x^* - x)^T$$

$$= E(\hat{x} + CY - x)(\hat{x} + CY - x)^T$$

$$= E[(\hat{x} - x)(\hat{x} - x)^T] + E[CY(\hat{x} - x)^T]$$

$$+ E[(\hat{x} - x)\ Y^TC^T] + E[CYY^TC^T]$$

or

$$C(x^*, x^{*T}) = C(\hat{x}, \hat{x}^T) + E[C(Hx + v)(\hat{x} - x)^T]$$

$$+ E[(\hat{x} - x)(Hx + v)^TC^T] + E[CYY^TC^T]$$

Consider the term

$$E[C(Hx + v)(\hat{x} - x)^T] = E[(CHx(\hat{x} - x)^T] + E[C\sigma(\hat{x} - x)^T]$$

$$+ E\{Cv(H^TR^{-1}H)^{-1}H^TR^{-1}Y\}$$

$$= E\{Cv[(H^TR^{-1}H)^{-1}H^TR^{-1}(Hx + v)]^T\}$$

$$= E\{CV\ x\ H^TR^{-1}H(H^TR^{-1}H)\}$$

$$+ E\ Cvv^TR^{-1}H(H^TR^{-1}H)^{-1}$$

$$= 0 + CR\ R^{-1}H(H^TR^{-1}H)^{-1}$$

$$= CH(H^TR^{-1}H)^{-1} = 0$$

It follows then that

$$C(x^*,x^{*T}) = (HR^{-1}H)^{-1} + CRC^T.$$

In order to minimize the variance of the elements of the vector we minimize the diagonal elements of CRC^T, a positive semi-definite matrix. That is, we require that the diagonal elements of CRC^T be zero. But in order for the diagonal elements to be zero and CRC^T to be positive semi-definite, $CRC^T = 0$. But R is nonsingular, hence C must be the null matrix.

COROLLARY 2.1.1 If R in (2.1) is simply $\sigma^2 I$, then (2.1) reduces to

$$\hat{x} = (H^T H)^{-1} H^T Y \tag{2.4}$$

the least square estimator for x, whose covariance matrix is

$$(H^T H)^{-1}. \tag{2.5}$$

The proof is given by replacing R by $\sigma^2 I$ and noting that the solution minimizes the error sum of squares,

$$e^T e = (Y - Hx)^T (Y - Hx)$$

$$= \sum_{i=1}^{n} e_i^2$$

where $e = \{e_i\}$.

By using the properties of the pseudo inverse of a matrix an easy extension of the Gauss-Markov theorem is possible. In order to obtain a yet more generalization of the theorem involves rather complicated range space arguments hence that generalization is given in Chapter III.

THEOREM 2.2 Consider the linear model described by the vector equation

$$\underset{p\times 1}{y} = \underset{p\times n}{H}\ \underset{n\times 1}{x} + \underset{p\times 1}{v}$$

where $E(v) = \phi$ and $E(vv^T) = R$ is positive definite. The minimum variance linear estimate $\hat{x}$ of x such that $E(\hat{x}) = x$ whenever x is in the range space of H^T is given by

(I) For rank $H = n \leq p$

$$\hat{x} = (H^T R^{-1} H) + H^T R^{-1} y = (H^T R^{-1} H)^{-1} H^T R^{-1} y$$

$$R_{\hat{x}} = (H^T R^{-1} H) + = (H^T R^{-1} H)^{-1}$$

(II) For rank $H = p < n$

$$\hat{x} = (H^T R^{-1} H^+ H^T R^{-1} y = H^+ y$$

$$R_{\hat{x}} = (H^T R^{-1} H)^+ H^+ R H^{+T}$$

Proof: We require that $\hat{x} = By$ and $E(\hat{x}) = x$ whenever x is in the range space of H. These requirements imply that $E(\hat{x}) = BHx$, which implies that for x in the range space of H^T,

$$H^+ Hx = BHx = x$$

so that $BH = H^+ H$ on the range space of H^+. Moreover, $||E(x) - \hat{x}||$ is minimum for x in the range space of H^T. The covariance matrix of $R_{\hat{x}}$ of the estimate $\hat{x}$ is given by $R_{\hat{x}} = BRB^T$ and must be minimized subject to the constraint $BH = H^+ H$. To do this we adjoin constraint $BH = H^+ H$ to BRB^T using a matrix Lagrange multiplier λ and find conditions necessary to minimize

$$Q = BRB^T + \lambda^T [H^+ H - H^T B^T] + [H^+ H - BH]\lambda.$$

Employing the variational technique we obtain the first variation δQ

$$\delta Q = \delta B[RB^T - H\lambda] + [BR - \lambda^T H^T]\delta B^T.$$

Since δB is arbitrary, we find that setting $\delta Q = \phi$ implies

$$BR - \lambda^T H^T = \phi$$

or

$$B = \lambda^T H^T R^{-1}.$$

Multiplying the latter by H we obtain

$$H^+H = \lambda^T H^T R^{-1} H$$

so that using (Theorem 1.10) and setting $H^T R^{-1} H = M$ we have

$$\lambda^T = H + HM^+ + Y[I - MM^+]$$

where Y is arbitrary to within having the dimension of λ^T.

In case (I), rank $H = n \leq p$ so that M is nonsingular. Moreover, (Lemma 1.8) implies $H^+H = I$ so that

$$\lambda^T = M^{-1} = M^+$$

$$B = M + H^T R^{-1} = M^{-1} H^T R^{-1}$$

and

$$x = (H^T R^{-1} H)^+ H\, R \quad y = (H^T R^{-1} H)^+ H^T R^{-1} y$$

$$R = (H^T R^{-1} H)^+ = (H^T R^{-1} H)$$

This completes the proof of case (I).

In case (II), rank $H = p \leq n$ so that by (Lemma 1.8) we have $H^{T+}H^{T} = I$. Applying (Theorem 1.12) with $A = H^{T}R^{-1}$ and $C = H$ we have

$$C_1 = A^{+}AX = (H^{T}R^{-1})^{+}H^{T}R^{-1}H$$

and

$$A_1 = AC_1C_1^{+} .$$

It is easy to see by direct substitution into the four defining equations for $(H^{T}R^{-1})^{+}$ that $(H^{T}R^{-1})^{+} = RH^{+}$ and hence that

$$C_1 = RH^{T+}H^{T}R^{-1}H = H$$

$$A_1 = H^{T}R^{-1}HH^{+} = H^{T}R^{-1}.$$

Finally we have

$$(H^{T}R^{-1}H)^{+} = C_1^{+}A_1^{+} = H^{+}RH^{T+}$$

and the estimate $\hat{x}$ of $\hat{x}$ is

$$\hat{x} = \{(H^{T}R^{-1}H)^{+} + Y[I - (H^{T}R^{-1}H)(H^{+}RH^{T+})]\}H^{T}R^{-1}y$$

or

$$\hat{x} = (H^{T}R^{-1}H) = H^{T}R^{-1}y + Y\phi = H^{+}y.$$

This completes the proof of case (II).

2.2 The Recursive Form of the Estimator

In real time estimation problems (filtering [4], [5], [6]) it is necessary that the estimator be written in a recursive form. This can be done easily by the following formulation:

$$\hat{x}_N = \hat{X}_{N-1} - \hat{x}_{n-1} + [H_N V_N^{-1} H_N]^{-1} H_N V_N^{-1} Y_N$$

where $\hat{x}_N$ is the best estimate given N data vectors y_i (pxl) where

$$Y_N = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \text{a Npxl}$$

vector of observations

$$H_N = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} \qquad e_N = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

is Npxl mapping matrix relating Y_N, X the parameter vector and the Npxl error vector e_N in the linear model

$$Y_N = H_N X + e_N.$$

Assuming that the covariance matrix of e_N is block diagonal, that is

$$V_N = \mathrm{Cov}(e_N) = \begin{bmatrix} V_{11} & \phi & \cdots & \phi \\ \phi & V_{22} & \cdots & \phi \\ \vdots & \vdots & \vdots & \vdots \\ \phi & \phi & \cdots & V_{NN} \end{bmatrix}$$

then

$$\hat{X}_N = \hat{X}_{N-1} - \hat{X}_{N-1} + [H_{N-1}^T V_{N-1}^{-1} H_{N-1} + h_N^T V_{NN}^{-1} h_N]^{-1} [H_{N-1}^T V_{N-1}^{-1} Y_{N-1} + h_N^T V_{NN}^{-1} Y_N]$$

or

$$\hat{X}_N = \hat{X}_{N-1} + [H_{N-1}^T V_{N-1}^{-1} H_{N-1} + h_N^T V_{NN}^{-1} h_N]^{-1} [H_{N-1}^T V_{N-1}^{-1} Y_{N-1} + h_N^T V_{NN}^{-1} Y_N - H_{N-1}^T V_{N-1}^{-1} H_{N-1} \hat{X}_{N-1} - h_N^T V_{NN}^{-1} h_N \hat{X}_{N-1}]$$

which reduces to

$$\hat{X}_N = \hat{X}_{N-1} + [H_{N-1}^T V_{N-1}^{-1} H_{N-1} + h_N^T V_{NN}^{-1} h_N]^{-1} h^T V^{-1} [y_N - h_N \hat{X}_{N-1}]$$

since

$$\hat{X}_{N-1} = [H_{N-1}^T V_{N-1}^{-1} H_N]^{-1} H_{N-1}^T V_{N-1}^{-1} Y_{N-1}.$$

A recursive form for the $\mathrm{Cov}\ \hat{X}_N$ can be obtained using the "inside-out" rule for inverting the sum of a positive-semi definite matrix and a

positive definite matrix (See page 25). That is,

$$\text{Cov } \hat{X}_N = [H_N^T V_N^{-1} H_N]^{-1}$$

$$= [H_{N-1}^T V_{N-1}^{-1} H_{N-1} + h_N^T V_{NN}^{-1} h_N]^{-1}$$

$$= (H_{N-1}^T V_{N-1}^{-1} H_{N-1})^{-1}$$

$$(H_{N-1}^T V_{N-1}^{-1} H_{N-1})^{-1} h_N^T [V_{NN} + h_N (H_{N-1}^T V_{N-1}^{-1} H_{N-1})^{-1} h_N]^{-1} h_N$$

$$(H_N^T V_{N-1}^{-1} H_{N-1})^{-1} = \text{Cov } (X_{N-1})$$

$$- \text{Cov}(X_{N-1}) h_N^T [V_{NN} + h_N \text{Cov}(\hat{X}_{N-1}) h_N^T]^{-1} h_N \text{ Cov}(\hat{X}_{N-1}),$$

which gives a recursive way for computing $\text{Cov}(\hat{X}_N)$ as a function of the $\text{Cov}(\hat{X}_{N-1})$.

This technique allows one to invert large covariance matrices by simply inverting a smaller dimensional matrix.

2.3 The Gauss-Markov Theorem When the Parameter Vector is Random

In section 2.1 the parameter vector x was assumed to be a constant vector. There are instances when it is natural to assume that the vector x in the linear model

$$y = hx + v \tag{2.6}$$

is random with the following statistical properties

$$E(x) = \mu_x \tag{2.7}$$

$$E(x-\mu_x)(x-\mu_x) = R_{xx} \tag{2.8}$$

The vector v is a $p \times 1$ vector of random elements such that

(2.9) $$E(v) = \phi$$

(2.10) $$E(vv^T) = R$$

(2.11) $$E(vx^T) = \phi$$

From (2.7) - (2.11) it follows easily that

(2.12) $$E(y) = h\mu_x$$

$$R_{yy} = E(y - h\mu_x)(y - h\mu_x - h\mu_x)^T$$

$$= E(hx + v - h\mu_x)(hx + v - h\mu_x)^T$$

$$= E[h(x-\mu_x)(x-\mu_x)^T h^T + v(x-\mu_x)^T h^T + h(x-\mu_x)v^T + vv^T]$$

(2.13) $$R_{yy} = h\, R_{xx}\, h^T = R$$

Also $R_{xy} = E[x-\mu_x)(y-\mu_y)^T] = E[(x-\mu_x)(x-\mu_x)^T h^T + v^T] = R_{xx}h^T$.

We will consider the class of estimators defined by the formula

(2.14) $$\hat{x} = a + Ay$$

where a is a vector of real numbers, A a matrix defined on the real numbers selected so that

(2.15) $$E[\hat{x} - x] = \phi$$

and

(2.16) $$Q = E[(\hat{x} - x)(\hat{x} - x)^T]$$

is minimized in the usual sense.

Consider the constraint (2.15)

$$\phi \equiv E(\hat{x} - x) = E[a + Ay - x]$$

or

$$\phi = a + Ah\mu_x - \mu_x$$

Two cases are immediate. These are

(1) μ_x is known

(2) μ_x is unknown.

THEOREM 2.3 Let $a + Ay$ be a linear estimator of x in the linear model (2.6). Then the optimum values of a and A for which

$$E[a + Ay - x][a + Ay - x]^T$$

is a minimum are

(2.16) $$a^* = \mu_x - A^* h\mu_x$$

(2.17) $$A^* = R_{xx}h^T[hR_{xx}h^T + R]^{-1}$$

The variance of the estimator is

$$R_{xx}h^T[h^TR_{xx}h + R]^{-1}hR_{xx}.$$

Proof: Let $\hat{x} = a + Ay$ be an estimator for x. Let μ_x be known and

$$Q = E[(\hat{x} - x)(\hat{x} - x)^T]$$

$$= E[(a + ay - x)(a + Ay - x)^T]$$

$$= aa^T + a\mu_y A^T - a\mu_x^T + A\mu_y a^T + A[R_{yy} + \mu_y\mu_y^T]A^T$$

$$- A[R_{yx} + \mu_y\mu_x^T] - \mu_x a^T - [R_{xy} + \mu_x\mu_y^T]A^T + [R_{xx} + \mu_x\mu_x^T].$$

A necessary condition for Q to be minimal is for the first variations in Q with respect to a and the first variation in Q with respect to A to be simultaneously the null vector and the null matrix, respectively. Let $\delta_a Q$ and $\delta_A Q$ denote these variations. Hence

$$\delta_a Q = a[a^T + \mu_y^T A^T - \mu_x] + [a = A\mu_y - \mu_x]\delta a^T$$

$$\delta_A Q = \delta A[\mu_y a^T + (R_{yy} + \mu_y\mu_y^T)A^T - (R_{yx} + \mu_y\mu_x^T)]$$

$$[a\mu_y + A(R_{yy} + \mu_y\mu_y^T) - (R_{xy} + \mu_x\mu_y^T)]\delta A^T.$$

The constraints $\delta_a Q \equiv \phi$ and $\delta_A Q \equiv \phi$ for all δa and δA respectively implies that

$$a^* + A^*\mu_y - \mu_x \equiv \phi$$

and $a\mu_y + A(R_{yy} + \mu_y\mu_y^T) - (R_{xy} + \mu_x\mu_y^T)$. The first condition implies $a^* = \mu_x - A^*\mu_y$. Since $\mu_y = h\mu_x$, it follows that

$$a^* = \mu_x - A^* h\mu_x.$$

The second condition implies that

$$A^* = R_{xy}R_{xx}^{-1} = R_{xx}h^T[hR_{xx}h^T + R]^{-1}$$

the desired results. We note that

$$\hat{x} = \mu_x + R_{xy}R_{yy}^{-1}(y - h\mu_x). \tag{2.18}$$

The covariance matrix, $R_{\hat{x}}$, for $\hat{x}$ follows easily from the definition

$$\begin{aligned} R_{\hat{x}} &= R_{xy}R_{yy}^{-1}R_{yy}R_{yy}^{-1}R_{yx} \\ &= R_{xy}R_{yy}^{-1}R_{yx} \\ &= R_{xx}h^T[h^TR_{xx}^{-1}h + R]^{-1}h\,R_{xx}. \end{aligned}$$

The proof is complete.

It is clear from (2.16) or (2.18) that if μ_x is not known then the estimate (2.18) is not computable. Consider the case for μ_x is not known.

THEOREM 2.4 The optimal values of a and A for which

(i) $E[a + Ay - x] = 0$ for all values of μ_x

(ii) $E[(a + Ay - x)(a - Ay - x)^T$ is a minimum

are

$$a^* = 0$$

$$A = (h^TR^{-1}h)^{-1}h^TR^{-1}.$$

The covariance matrix of the estimator is

$$(h^TR^{-1}h)^{-1}.$$

Proof: The condition (i) implies that

$$a + Ah\,\mu_x - \mu_x = 0$$

for all μ_x. This in turn implies that

$$a + (Ah - I)\mu_x \equiv 0$$

for all μ_x. Hence

$$a = \phi$$

$$Ah - I = \phi$$

or we select A so that

$$Ah = I.$$

Let

$$Q = E[Ay - x)(Ay - x)^T] + \lambda^T[I - h^TA^T] + [I - Ah]\lambda \tag{2.19}$$

where λ is a vector of Lagrangian multipliers. Substituting $Ah = I$ and $y = hx + v$ into (2.19) we get

$$Q = ARA^T + \lambda^T[I - h^TA^T] + [I - Ah]\lambda.$$

Equating the first variation of Q with respect to A to zero one obtains

$$AR - \lambda^T h^T = \phi \tag{2.20}$$

$$AR = \lambda^T h^T$$

$$A = \lambda^T h^T R^{-1} \tag{2.21}$$

Multiplying both sides of (2.21) by h on the right

$$I = \lambda^T h^T R^{-1} h.$$

From (2.21), it follows that

$$A = (h^T R^{-1} h)^{-1} h^T R^{-1}.$$

2.4 On Estimating a Subvector of x

The minimum variance linear unbiased estimate for x in the linear model

$$y = hx + v$$

is by Theorem 2.1

$$\hat{x} = (h^T R^{-1} h)^{-1} h^T R^{-1} y \tag{2.22}$$

Let x be partitioned such that

$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix}$$

where $x^{(1)}$ is $q \times 1$ vector and $x^{(2)}$ is $n-q \times 1$ vector. Suppose we wish to estimate $x^{(1)}$ by a minimum variance estimator and yet not estimate the total vector x by using (2.22).

THEOREM 2.5 The minimum variance linear unbiased estimator for $x^{(1)}$, the first q elements of $\hat{x}$ is given by the vector (2.24).

Proof: By partitioning the covariance matrix of $\hat{x}$ we can write the covariance matrix $R_{\hat{x}}$ of $\hat{x}^{(1)}$.

$$(2.23)\quad R_{\hat{x}} + [(h_1 h_2)^T R^{-1}(h\ h\ )]^{-1} = \begin{bmatrix} h_1^T R^{-1} h_1 & h_1^T R^{-1} h_1 \\ h_2^T R^{-1} h_1 & h_2^T R^{-1} h_2 \end{bmatrix}$$

Inverting the matrix (2.23) we find that

$$R_{\hat{x}}(1) = [h_1 R^{-1} h_1 - h_1 R^{-1} h_1 (h_2^T R^{-1} h_2)^{-1} h_2^T R^{-1} h_1]^{-1}$$

$$= [h_1 (R^{-1} - R^{-1} h_1 \{h_2^T R^{-1} h_2\}^{-1} h_2^T R^{-1}) h_1]^{-1}$$

Following the form of (2.22) we write

$$(2.24)\quad \hat{x}^{(1)} = [h_1 (R^{-1} - R^{-1} h_1 \{h_2^T R^{-1} h_2\}^{-1} h_2^T R^{-1}) h_1]^{-1}$$

$$h_1^T (R^{-1} - R^{-1} h_1 \{h_2^T R^{-1} h_2\}^{-1} h_2^T R^{-1} y.$$

On taking the expectation of $x^{(1)}$

$$E\,\hat{x}^{(1)} = [h_1 (R^{-1} - R^{-1} h_1 \{h_2^T R^{-1} h_2\}^{-1} h_2 R^{-1} h_1]^{-1} \; x^{(1)}$$

$$h_1^T (R^{-1} - R^{-1} h_1 \{h_2^T R^{-1} h_2\}^{-1} h_2^T R^{-1} (h_1 h_2) \; x^{(2)}$$

$$= x^{(1)}$$

since

$$(R^{-1} - R^{-1} h_1 \{h_2^T R^{-1} h_2\}^{-1} h_2^T R^{-1} h_2\, x^{(2)} = \phi$$

The fact that $\hat{x}$ is minimum variance implies that indeed $\hat{x}^{(1)}$ is minimum variance.

This partitioning can help in eliminating the necessity of computing unwanted parameters which may represent systematic sources of errors.

2.5 References

[1] Graybill, F. A., An Introduction to Linear Statistical Model, Vol. 1, McGraw-Hill Co., 1961.

[2] Decell, H. P. and Odell, P. L., "A Note on a Generalization of the Gauss-Markov Theorem," Texas Journal of Science, March 1966, pp. 21-24.

[3] Lewis, T. O. and Odell, P. L., "A Generalization of Gauss-Markov Theorem, Journal of American Stat. Assoc., Vol. 61, 1966, pp. 1063-1066.

[4] Kalman, R. E., "A New Approach to Linear Filtering and Prediction problems, "Trans ASME, Series D, Journal of Basic Engineering 82, 1960, pp. 34-45.

[5] Bendat, J. S. and Piersol, A. G., Measurement and Analysis of Random Data, John Wiley and Sons, Inc., 1966.

[6] Blum, M., "Fixed Memory Least Squares Filters Using Recursive Methods," IRE Transactions of the Professional Group on Information Theory, Vol. IT-3, #3, Sept, 1957.

Chapter III

A GENERALIZATION OF THE GAUSS-MARKOV THEOREM

This chapter contains a generalization of the Gauss-Markov Theorem based on the properties of the generalized inverse of a matrix as defined by Penrose. A minimum variance vector estimate $\hat{x}$ of a parameter vector x is given for the linear model of less than full rank. Since linear unbiased estimates may not always exist for this case the unbiased constraint is replaced by the more general constraint that the norm

$$||E(\hat{x}) - x||$$

is minimized.

3.1 Introduction

Several authors [1], [2], [3], [4], [5], and [6] have considered least square and minimum variance estimation of parameters in a less than full rank linear model,

$$y = Hx + e. \tag{3.1}$$

In (1) y denotes a $p \times 1$ vector of observations; H a known real $p \times n$ matrix; x an $n \times 1$ vector of fixed but unknown parameters to be estimated; and e a vector of random errors such that $E\{e\} = \phi$ and $E\{ee^T\} = V$, a positive definite matrix, where E and ϕ denote the expectation operator and the null vector, respectively.

It is well-known that in the full-rank case (rank of H equal to n) that the Gauss-Markov theorem yields the minimum variance linear

unbiased estimator $\hat{x}$ where

$$\hat{x} = (H^T V^{-1} H)^{-1} H^T V^{-1} y. \tag{3.2}$$

Its covariance matrix $R_{\hat{x}}$ is simply

$$(H^T V^{-1} H)^{-1}. \tag{3.3}$$

In [6] Decell and Odell showed that if the rank of H is $p < n$, then

$$\hat{x} = H^+ y \tag{3.4}$$

and

$$R_{\hat{x}} = H^+ V H^{+T} \tag{3.5}$$

where the superscript $^+$ denotes the generalized inverse [7] and [4]. It is interesting to note that $\hat{x}$ in this case is also the least square estimate of x.

For practical reasons most investigators studied the problem of minimum variance estimation for those parameters or those linear functions of parameters which are linearly estimable [8]. It is our purpose in this chapter to formulate a generalization of the Gauss-Markov theorem which will include the results (2) and (4) as special cases and discuss briefly the meaning that can be attached to the estimator.

3.2 Notation and Preliminaries

We seek a linear, minimum variance, unbiased estimator $\hat{x}$ of x, if such an estimator exists. That is, if $\hat{x} = By$, we are to find a real

matrix B such that $E\hat{x} = x$ and $R_{\hat{x}} = E[(\hat{x} - x)(\hat{x} - x)^T]$ is minimum in the sense that if z is any other linear unbiased estimate of x and $V_z = E[(z - x)(z - x)^T]$, then $q^T[V_z - V_{\hat{x}}]\, q > 0$ for any $p \times 1$ vector $q = \phi$. These conditions imply that $E(\hat{x}) = HBx = x$ so that $BH = I$, where I is the $n \times n$ identity matrix.

If the rank of H is $p < n$ we cannot require that $E(\hat{x}) = x$, since in this case H has no left inverse. We can, however, modify this requirement by requiring the norm $||E(\hat{x}) - x||$ be minimum and then in turn select from this class of linear estimators one which has minimum variance with respect to the range space of H^T. The norm used here is $||E(\hat{x})-x|| = [(E(\hat{x})-x)^T\ (E(\hat{x}) - x)]^{\frac{1}{2}}$. Such an estimator will be called a best linear estimator.

To facilitate reading, we list again some properties of the Penrose [7], [4], [9] pseudo-inverse used in obtaining this result.

P1) For every matrix A there exists a unique matrix A^+ such that

$$A A^+ A = A$$
$$A^+ AA^+ = A^+$$
$$(A^+A)^T = A^+A$$
$$(AA^+)^T = AA^+.$$

We call A^+ the pseudo-inverse of A.

P2) $(AC)^+ = C_1^+ A_1^+$ where $AC = A_1C_1$, $C_1 = A^+AC$, and $A_1 = AC\ C_1^+$

P3) $(A^+)^T = (A^T)^+$

P4) All solutions of the matrix equation $AXB = C$ are given by $X = A^+CB^+ + y - A^+AYBB^+$ if and only if $AA^+CB^+B = C$ where Y has the dimension of X.

P5) Range of A^T equals the range of A^+, that is $R(A^T) = R(A^+)$. A^+A and AA^+ are, respectively, the projection operators on the range spaces of A^+ and A.

P6) For any $n \times n$ matrix A and vector z, $z = z_1 + a_2$ where $z_1 \epsilon R(A^+)$, $a_2 \epsilon N(A)$, and z_1 is orthogonal to z_2.

3.3 The Main Result

We are now ready to establish a generalization of the Gauss-Markov Theorem.

THEOREM 3.1 Consider the linear model described by the vector equation

$$\underset{p\times 1}{y} = \underset{p\times n}{H}\ \underset{n\times 1}{x} + \underset{p\times 1}{e}$$

where $E(e) = \phi$ and $E(ee^T) = V$ is positive definite. The best linear estimator $\hat{x}$ of x is given by:

$$\hat{x} = M^+H^TV^{-1}y$$

and

$$V_{\hat{x}} = M^+$$

where

$$M = H^TV^{-1}H.$$

Proof: We require that $\hat{x} = By$ and $E(\hat{x}) = x$ whenever $x \epsilon R(H^T)$. These requirements imply that $E(\hat{x}) = BHx$ and (P5) implies that for x in $R(H^T)$,

$$H^+Hx = BHx = x.$$

Let $x = x_1 + x_2$ where $x_1 \epsilon R(H^T)$, $x_2 \epsilon N(H)$.

Then

$$||E(\hat{x}) - x|| = ||BHx_1 + BHx_2 - x|| = ||BHx_1 - x|| = ||x_2||.$$

Thus it follows that $||E(\hat{x}) - x||$ is minimum with respect to $R(H^T)$ for $x \notin R(H^T)$. The covariance matrix $V_{\hat{x}}$ of the estimator $\hat{x}$ is given by $V_{\hat{x}} = BVB^T$ and must be minimized subject to the constraint $BH = H^+H$. To do this we adjoin the constraint $BH = H^+H$ to BVB^T using a matrix Lagrange multiplier λ and find conditions necessary to minimize

$$Q = BVB^T + \lambda^T [H^+H - H^TB^T] + [H^+H - BH]\lambda.$$

Employing the variational technique [3] we obtain the first variation δQ,

$$\delta Q = \delta B[VB^T - H\lambda] + [BV - \lambda^T H^T]\ \delta B^T.$$

Since δB is arbitrary, we find that setting $\delta Q = \phi$ implies

$$BV - \lambda^T H^T = \phi$$

or

$$B = \lambda^T H^T V^{-1}.$$

Multiplying the latter by H we obtain

$$H^+H = \lambda^T H^T V^{-1} H$$

so that using (P4) and setting $H^TV^{-1}H = M$ we have

$$\lambda^T = M^+ + y(I - MM^+)$$

where y is arbitrary to within having the dimension of λ^T. To see this we need to show that

$$H^+HM^+M = H^+H.$$

$$H^+H(H^TV^{-1}H)^+(H^TV^{-1}H) = H^+H.$$

It follows that $R(M^+) = R(M^T) = R(M)$. Hence we must show that $R(M) = R(H^T)$. We observe that $R(H^T) \supset R(M)$ and $N(H) \subset N(M)$. Suppose there exists $x \in N(M)$ such that $x \notin N(H)$. Then $Hx \neq 0$ and $Mx = 0$. But since V^{-1} is positive definite $x^TMx \neq 0$ which implies $Mx \neq 0$. This is a contradiction. Hence $N(H) = N(M)$. Since $R(M^T)$ and $N(M)$ are orthogonal spaces and their direct sum is the n-dimensional vector space x_n, it follows that $R(M) = R(H^T)$. We now observe that the columns of H^+H are in $R(M^+)$ thus $M^+MH^+H = H^+H$. Taking the transpose of both sides gives

$$H^+HM^+M = H^+H.$$

Assume that the rank of H is $q \leq \min(n,p)$. Then

$$B = \lambda^T H^T V^{-1}$$

$$= (M^+ + y\,[I - MM^+])H^TV^{-1}$$

$$= M^+H^TV^{-1} + y\,[I - MM^+]\,H^TV^{-1}.$$

To establish the second term is ϕ, i.e. $[I - MM^+]H^TV^{-1} = \phi$,

we observe that $[I - MM^{+}]$ is an orthogonal projection on the null space of M^{+}. We need to show that $N(M^{+}) = N(H)$. Since $M = H^{T}V^{-1}H$, then it certainly follows that $N(M) = N(M^{T})$. Also note that $N(M) \supseteq N(H)$. Thus suppose there exists an $x \in N(M)$ such that $x \notin N(H)$. Hence, it follows $Hx \; R(H)$. Since V^{-1} is positive definite, then V^{-1} does not rotate Hx into the null space of H^{T}. Hence $H^{T}V^{-1}Hx \neq 0$, which implies $x \notin N(M)$. This is a contradiction. Thus $N(M) = N(H)$. Now $N(M) = N(M^{T}) = N(M^{+})$ which implies $N(M^{+}) = N(H)$ and consequently $(I - MM^{+})H^{T}V^{-1} = \phi$ since $R(H^{T}) = N(H)^{\perp}$, where $(\cdot)^{\perp}$ denotes the orthogonal complement of $(\cdot)$. Hence

$$\hat{x} = By = M^{+}H^{T}V^{-1}y$$

and the covariance matrix

$$R_{\hat{x}} = BVB^{T} = M^{+}MM^{+T} = M^{+}$$

the desired results.

It should be noted that if the rank of H is equal to $n \leq p$, then $H^{+}H = I$ and $\hat{x}$ reduces to (3.2) and its covariance matrix is given by (3.3). If the rank of H is $p \leq n$ then $HH^{+} = I$, $(H^{T}V^{-1})^{+} = VH^{T+}$ and (3.4) and (3.5) follow.

3.4 Comparison of Least Squares and Minimum Variance Estimates of Regression Parameters

It is of interest to compare the least squares estimate of the state vector to that of the minimum variance estimate of the state vector.

Magness and McGuire have been able to give an extensive analysis in comparing these two estimates whenever the regression matrix of the linear model is of full-rank (columns linearly independent). They were able to establish the inequality

$$V_{LS} \leq \frac{1}{4}(\lambda_{max} + \lambda_{min})\left(\frac{1}{\lambda_{max}} + \frac{1}{\lambda_{min}}\right)V_{MV}$$

where V_{LS} and V_{MV} are the covariance matrices of the least squares estimate and minimum variance estimate, respectively. λ_{max} and λ_{min} are the maximum and minimum eigen values of the correlation matrix ρ of the error vector. The above inequality places an upper bound on how much is lost by use of the least squares estimate of the state vector to that of the minimum variance estimate of the state vector.

In the following theorem it will be shown that the least squares estimate of the state vector will have the same covariance matrix as that of the mean-square-error estimate of the state vector, whenever the regression matrix of the linear model has all of its rows linearly independent.

THEOREM 3.2 Consider the linear model described by the vector equation

$$\underset{p\times 1}{y} = \underset{p\times n}{H}\ \underset{n\times 1}{X} + \underset{p\times 1}{e}$$

where $E(e) = \phi$, $E(ee^T) = V$ is positive definite, $R(H) = p$. Then the covariance matrix of the least-squares estimate of the state vector equals the covariance matrix of the mean-square-error estimate of the state vector.

Proof: The least squares estimate of the state vector is

$$\hat{x}_{LS} = (H^T H)^+ H^T y$$

$$= H^+ y.$$

The corresponding covariance matrix is

$$V_{LS} = H^+ V H^{T+}.$$

The mean-square-error estimate of the state vector is by Theorem 3.1

$$\hat{x} = H^+ y.$$

The corresponding covariance matrix is

$$V_{\hat{x}} = H^+ V H^{T+}.$$

Thus it can be seen there is no loss in using the least squares estimate whenever the rows of the regression matrix are linearly independent.

3.5 References

[1] Dwyer, P. S., "Generalizations of a Gaussian Theorem," Ann. of Math. Statis. 22 No. 1 (1958), pp. 106-117.

[2] Rao, C. R., "A Note on a Generalized Inverse of a Matrix with Application to Problems in Mathematical Statistics," Journal of the Royal Statistical Society, Series B, 24, (1962), pp. 152-158.

[3] Goldman, A. J. and Zelen, M., "Weak Generalized Inverse and Minimum Variance Linear Unbiased Estimation," Journal of Research of the National Bureau of Standards - B Mathematical and Mathematical Physics 68B, No. 4, (1964), pp. 151-172.

[4] Price, C. M., The Matrix Pseudoinverse and Minimal Variance Estimates, SIAM Review, 6, No. 22 (1964), pp. 115-120.

[5] Chipman, J. S., "On Least Squares with Insufficient Observations," Journal of the Amer. Statist. Assoc., 59, (1964), pp. 1078-1111.

[6] Decell, H. P., and Odell, P. L., "A Note Concerning a Generalization of the Gauss-Markov Theorem," Texas Journal of Science, 27, No. 1, (1966), pp. 21-24.

[7] Penrose, R., "A Generalized Inverse for Matrices," Proc. Cambridge Philos. Soc., 51, (1955), pp. 406-413.

[8] Graybill, Franklin A., An Introduction to Linear Statistical Models, Vol. I, pp. 227-228.

[9] Cline, R. E., "Note on the Generalized Inverse of the Product of Matrices," SIAM Review, 6, No. 1, (1964), pp.57-58.

[10] Magness, T. A. and McGuire, T. B., "Comparison of Least Squares and Minimum Variance Estimates on Regression," The Annal of Math.Statis., Vol. 33, No. 2, June, 1962.

Chapter IV

THE GAUSS-MARKOV THEOREM AND ITS RELATION TO CONTINUOUS RECURSIVE ESTIMATION

Kalman [1] noted that there existed a direct relationship between the so-called recursive estimators associated with dynamic linear models and the minimum variance unbiased estimator for a fixed[1] parameter vector related to an observation vector by the following linear model:

$$y(t) = H(t)\Phi(t,T)\, x(T) + v(t) \tag{4.1}$$

where $y(t)$ is an observed $p \times 1$ vector function of the real variable t for $t_0 \leq t \leq T$. Also, $H(t)$ and $\Phi(t,T)$ are known $p \times n$ and $n \times m$ matrices whose elements are real valued functions of t; $x(T)$ is an $n \times 1$ fixed vector of unknown parameters we wish to estimate; and $v(t)$ is a random $p \times 1$ vector of random real valued functions of t such that,

$$E\{v(t)\} = 0 \qquad \text{for } t_0 \leq t \leq T \tag{4.2}$$

and

$$E\{v(t)v(s)^T\} = R(t)\delta(t,s) \tag{4.3}$$

where $\delta(t,s) = 1, 0$ if $t = s$ or $t \neq s$, respectively. The symbols $E\{\cdot\}$ and $\{\cdot\}$ denote the expected value and the transpose of $\{\cdot\}$, respectively.

---

1 If the parameter vector $x(T)$ is random and independent of $v(t)$ for all $t_0 \leq t \leq T$ and $Ex(T)$ is not known it can be shown that the estimate defined by (4.6) is still best.

4.1 The Gauss-Markov Theorem for Continuous Data

The following definitions and theorems are basis for the logical development of the relationship:

DEFINITION 4.1 By a linear estimator $\hat{\Pi}$ (for continuous data $Y(t)$ of a linear combination of the elements of $x(T)$, say

$$\Pi = p^T x(T) \tag{4.4}$$

we mean

$$\hat{\Pi} = \int_{t_0}^{T} w^T(t) y(t)\, dt \tag{4.5}$$

where p is an arbitrary but known constant vector, and $w(t)$ is an arbitrary (at least piecewise continuous) vector function of the real variable t.

It is clear that $\hat{\Pi}$ is a random variable, since it is a function of the observations $Y(t)$. The Gauss-Markov theorem notes that by selecting $w(t)$ properly one obtains a minimum variance unbiased linear estimator for Π which yields a minimum variance unbiased estimator for $x(T)$ if the vector p is chosen as the unit vector which forms a basis for the n-dimensional Euclidean space.

Consider the following Gauss-Markov theorem for continuous data:

THEOREM 4.1 The minimum variance unbiased linear estimator $x(T)$ in the linear model defined by (1) is $\hat{x}(T)$ where

$$\hat{x}(T) = M^{-1}(t_0,T) \int_{t_0}^{T} \Phi^T(t,T)\, H^T(t)\, R^{-1}(t)\, y(t)\, dt \tag{4.6}$$

$$M(t_0T) = \int_{t_0}^{T} \Phi^T(t,T)\, H^T(t)\, R^{-1}(t)\, H(t)\, \Phi(t,T)\, dt \tag{4.7}$$

and in (4.3)

$$R(t,s) = R(t)\delta(t,s) \text{ where } \delta(t,s) = 1 \text{ for } t = s, \ = 0 \text{ for } t \neq s.$$

Proof: In (4.5) let

$$w(t) = R^{-1}(t)\, H(t)\, \Phi(t,T)\, M^{-1}(t_0T)p.$$

Since $M^{-1}(t_0,T)$ is symmetric it is straight forward using (4.7) to show that $\hat{\Pi}$ is an unbiased estimator for Π, that is $E\{\hat{\Pi}\} = \Pi$.

Let $\overset{*}{\Pi} \neq \hat{\Pi}$ be any other unbiased estimator for Π, then we can write

$$\overset{*}{\Pi} = \int_{t_0}^{T} [w(t) + r(t)]^T\, y(t)\, dt. \tag{4.8}$$

Since $E\overset{*}{\Pi} = \Pi$, it follows that

$$\int_{t_0}^{T} r(t)\, H(t)\, \Phi(t,T)\, dt = 0 \tag{4.9}$$

Consider the variance of $\overset{*}{\Pi}$, that is

$$\text{Var}\,(\overset{*}{\Pi}) = E\left[\int_{t_0}^{T} [w(t) + r(t)]^T\, v(t)dt\right]^2$$

$$= E\left\{\int_{t_0}^{T}\int_{t_0}^{T} [s(t) + r(t)]^T\, v(t)\, v^T(s)\, [w(s) + r(s)]^T\, dtds\right\}$$

$$= \int_{t_0}^{T} w^T(t)\, R(t)\, w(t)\, dt + 2\int_{t_0}^{T} r^T(t)\, R(t)\, w(t)\, dt$$

$$+ \int_{t_0}^{T} r^T(t)\, R(t)\, r(t)\, dt.$$

Note that

$$\mathrm{Var}(\hat{\Pi}) = \int_{t_0}^{T} w^T(t)\, R(t)\, w(t)\, dt,$$

$$\int_{t_0}^{T} r(t)\, R(t)\, r(t)\, dt \geq 0$$

since $R(t)$ is positive definite for all t in the interval $[t_0,T]$ and by (4.9) and the definition of $w(t)$

$$\int_{t_0}^{T} r^T(t)\, R(t)\, w(t)\, dt = \int_{t_0}^{T} r^T(t)\, R(t)\, R^{-1}(t)\, H(t)\, \Phi(t,T)\, dt$$

$$M^{-1}(t_0,T)p$$

$$= 0.$$

Hence,

$$\mathrm{Var}\,(\overset{*}{\Pi}) \geq \mathrm{Var}\,\hat{\Pi}$$

for all $r(t)$, and equality holds for $r(t) \equiv 0$.

If $p = e_i$, $i = 1, \ldots, n$ the unit matrix with unity in the i^{th} position, then the minimum variance linear unbiased estimator for $X(T)$ follows and is given by (4.6).

4.2 A Dynamic Model

In application associated with linear dynamic filters the "so-called" state transition matrix $\Phi(t,T)$ is obtained from a system of linear homogeneous differential equations which usually appear in the form

$$\frac{dx(t)}{dt} = f(t)x(t) + G(t)u(t) \tag{4.10}$$

where the vector $u(t)$ may be one of the following

1) $u(t) = 0$

2) $u(t)$ is a continuous time series from a random process such that

$$E\{u(t)\} = 0 \qquad \text{for all } t \tag{4.11}$$

$$E\{u(t)u^T(s)\} = Q(t,s). \tag{4.12}$$

Usually $Q(t,s)$ is assumed to be

$$Q(t,s) = Q(t)\ \delta(t,s) \tag{4.13}$$

where

$$\delta(t,s) = \begin{cases} 1 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases} \tag{4.14}$$

the so-called "white-noise" case.

For given $x(t)$, t, T the solution of (10) is well-known

$$x(T) = \psi(T,t)x(t) + \int_t^T \psi(T,s)\ G(s)\ u(s)\ ds \tag{4.15}$$

where $\psi(T,t)$ is a non-singular matrix such that

$$\text{(i)} \quad \psi(T,t) = \psi(T,t_1)\ \psi(t_1,t) \qquad t \leq t_1 \leq T.$$

$$\text{(ii)} \quad \frac{d\psi(T,t)}{dt} = \psi(T,t)$$

Generally one knows that

$$y(t) = H(t)\ x(t) + v(t) \tag{4.16}$$

and from (4.15) one obtains the transition matrix $\Phi(t,T)$, that is

$$x(t) = \psi^{-1}(T,t)\ x(T) + \psi^{-1}(T,t) \int_t^T \psi(T,s)\ G(s)\ u(s)\ ds \tag{4.17}$$

We shall consider here the case where $u(s) = 0$, and later the other case, that is

$$x(t) = \Phi(T,t)\ x(T)$$

where

$$\Phi(t,T) = \psi^{-1}(T,t). \tag{4.18}$$

Note that from (i) that

$$\psi^{-1}(T,t) = \psi^{-1}(t_1,t)\ \psi^{-1}(T,t_1)$$

or

$$\Phi(t,T) = \Phi(t,t_1)\ \Phi(t_1,T). \tag{4.19}$$

Substituting (4.18) into (4.16) one obtains the linear model defined by (4.1) and Theorem 4.1 applies.

4.3 The Recursive Form of $\hat{x}(T)$.

Using the linearity properties of the integral operator and (4.19) one can give a recursive form of the estimator $\hat{x}(T)$. Let t_1 be such that $t_0 < t_1 < T$. Then

$$\hat{x}(T) = M^{-1}(t_0T) \int_{t_0}^{t_1} \Phi^T(t_1,T)\Phi^T(t,t_1)H^T(t)R^{-1}(t)y(t)dt$$

$$+ \int_{t_1}^{T} \Phi^T(t,T)H^T(t)R^{-1}(t)y(t)dt$$

Also

$$M(t_0T) = \int_{t_0}^{t_1} \Phi^T(t_1,T)\Phi^T(t,t_1)H^T(t)R^{-1}(t)H(t)\Phi(t,t_1)\Phi(t_1,T)\, dt$$

$$+ \int_{t_1}^{T} \Phi^T(t,T)H^T(t)R^{-1}(t)H(t)\Phi(t,T)\, dt$$

$$= \Phi^T(t_1,T)M(t_0,t_1)\Phi(t_1T) + M(t_1,T).$$

Hence

$$\hat{x}(t) = [\Phi^T(t_1,T)M(t_0,t_1)\Phi(t_1,T) + M(t_1,T)]^{-1} \tag{4.20}$$

$$[\Phi^T(t_1,T)M(t_0,t_1)\hat{x}(t_1) + M(t_1,T)\overset{*}{x}(T)]$$

where

$$\overset{*}{x}(T) = M(t_1,T) \int_{t_1}^{T} \Phi^T(t,T)H^T(t)R^{-1}(t)y(t)\ dt.$$

Note that by (4.3) that $\overset{*}{x}(T)$ and $\hat{x}(t_1)$ are independent, and (4.20) is the familiar formula for combining independent unbiased estimators of $x(T)$.

Let

$$\bar{x}(T) = \Phi^{-1}(t_1,T)\hat{x}(t_1)$$

and $\overset{*}{x}(t)$ be the unbiased estimators of $x(T)$ since

$$E\bar{x}(T) = \Phi^{-1}(t_1,T)\ x(t_1) = x(T)$$

and

$$E\overset{*}{x}(T) = x(T).$$

Also we note that

$$\begin{aligned} \text{Var}\ \bar{x}(T) &= \Phi^{-1}(t_1,T)\ M^{-1}(t_0,t_1)\ \Phi^{-T}(t_1,T) \\ &= [\Phi^T(t_1,T)\ M(t_0,)\ \Phi(t_1,T)]^{-1} \\ &= \overline{M(t_0,T)}^{-1} \end{aligned}$$

and

$$\text{Var}\ \overset{*}{x}(T) = M^{-1}(t_1,T).$$

Hence (4.20) can be written as

$$\hat{x}(T) = [\overline{M(t_0,T)} + M(t_1,T)]^{-1} [\overline{M(t_0,T)}\ \bar{x}(T) + M(t_1,T)\ \hat{x}(T)]$$

$$= [(\text{Var}\ \bar{x}(T)^{-1} + (\text{Var}\ \overset{*}{x})^{-1}]^{-1} [(\text{Var}\ \bar{x}(T)^{-1}\ \bar{x}(T)$$

$$+ (\text{Var}\ \overset{*}{x}(T)^{-1}\overset{*}{x}(T)]$$

or equivalently

$$\hat{x}(T) = \text{Var}\ \overset{*}{x}(T)\,[\text{Var}\ \bar{x}(T) + \text{Var}\ \overset{*}{x}(T)]^{-1}\ \bar{x}(T)$$

$$+ \text{Var}\ \bar{x}(T)\,[\text{Var}\ \bar{x}(T) + \text{Var}\ \overset{*}{x}(T)]^{-1}\ \bar{x}(T),$$

a form of the Kalman Filter.

4.4 A Modification for Correlated Noise

Suppose that the assumption (4.3) is replaced by

$$R(s,t) = R(\tau)$$

where $\tau = |s - t|$. That is, the stochastic process from which $y(t)$ is a time series is covariance stationary.

We define (if the integral exists)

$$P_V(\omega) = \int_{-\infty}^{\infty} e^{i\omega\tau}\ R_V(\tau)\ d\tau$$

sometimes called the power density spectrum [2]. If

$$w(t) = \int_{-\infty}^{\infty} g(\tau)\ v\ (t - \tau)\ d\tau,$$

then

$$W(\omega) = G(\omega)\ V(\omega)$$

where $W(\omega)$, $G(\omega)$ and $V(\omega)$ are the Fourier transform of $w(t)$, $g(t)$, and $v(t)$, respectively. Also it is straight forward to show that

$$W(\omega)\ W^{*}(\omega) = G(\omega)\ G^{*}(\omega)\ V(\omega)\ V^{*}(\omega)$$

$$P_{W}(\omega) = |G(\omega)|^{2}\ P_{V}(\omega).$$

If we require $w(t)$ to be uncorrelated, that is,

$$E[w(t)] = \emptyset$$

and

$$Ew(t)w^{T}(t) = R_{W}(t)\ \delta(t,s)$$

then

$$P_{W}(\omega) = C$$

where C is a constant. Let $C = 1$ then

$$G(\omega) = [\frac{1}{P_{V}(\omega)}]^{1/2}$$

if $G(\omega)$ exists for all ω.

But $G(\omega) = F(g,t)$, the Fourier transform of g, hence

$$g(t) = F^{-1}(G,\omega).$$

the inverse Fourier transform of $G(\omega)$.

It is immediate if

$$|P_y(\omega)| > 0$$

for all ω then $G(\omega)$ is well defined. For those cases in which $g(t)$ is defined we simply compute

$$\bar{y}(t) = \int_{-\infty}^{\infty} g(\tau)\, y(t - \tau)\, d\tau$$

$$\bar{h}(t) = \int_{-\infty}^{\infty} g(\tau)\, h(t - \tau)\, d\tau$$

and consider the model

$$\bar{y}(t) = \bar{H}(t)\Phi(t,\tau)\, x(\tau) + \bar{u}(t)$$

where

$$\bar{u}(t) = \int_{-\infty}^{\infty} g(\tau)\, u(t - \tau)\, d\tau = \omega(t).$$

The conditions of Theorem 4.1 are valid and (4.6) can be used for estimating $\hat{x}(\tau)$ if $y(t)$ is replaced by $\bar{y}(t)$; $H(t)$ is replaced by $\bar{H}(t)$; and $R(t)$ replaced by $R_W(t)$.

4.5 References

[1] Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans ASME, Series D, Journal of Basic Engineering 82, 1960, pp. 34-45.

[2] Bendat, J. S. and Piersol, A. G., Measurement and Analysis of Random Data, John Wiley and Sons, Inc.

Chapter V

MATRIX LOWER BOUND FOR THE COVARIANCE MATRIX OF A VECTOR ESTIMATE

A matrix bound similar to the Cramer-Rao lower bound [1], [2], [3], for the covariance matrices of vector estimates can be formulated in a matrix notation which facilitates in the search for vector estimates with minimal covariance matrices. (We recall that one positive definite covariance matrix is by definition less than or equal to another provided the second minus the first is non-negative definite.) The derivation presented here is similar to that of Cramer's [4] in which he establishes the efficiency of vector estimates using the concept of ellipsoids of concentration.

A matrix lower bound for the covariance matrices of unbiased vector estimates of the unknown parameters in the linear regression model with correlated normal error is established. A direct result of this application is that the best linear estimate given by Gauss-Markov Theorem [5] still yields the minimum covariance matrix when compared with other vector estimates from the larger class composed of linear and non-linear vector estimates. This result has been noted [6] for the case of uncorrelated normal errors. When the variance is known, the result obtained here is similar to that given in [1] for estimates of various scalar functions of the parameters.

5.1 The Matrix Lower Bound

Let the joint density function of n random variables $Y = [Y_1, \ldots, Y_n]^T$ be

(5.1) $$L = L(y_1, \ldots, y_n;\ \beta_1, \ldots, \beta_p)$$

where $\beta = [\beta_1, \ldots, \beta_p]^T$ are unknown parameters which we wish to estimate. The symbol A^T denotes transpose of A; and the symbol E denotes the expectation operator. An arbitrary $n \times m$ matrix A will at times be denoted by $\{a_{ij}\}_n^m$, while the null matrix and the identity matrix will be denoted by Φ and I, respectively.

By definition of joint density function, it follows that

(5.2) $$1 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L \, dy_1 \ldots dy_n$$

It is assumed that the following regularity conditions hold:

(i) $$\frac{\partial}{\partial \beta_i} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L \, dy_1 \ldots dy_n \right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial L}{\partial \beta_i} \, dy_1 \ldots dy_n$$

$$i = 1, 2, \ldots, p.$$

(ii) $$\frac{\partial}{\partial \beta_j} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T_i L \, dy_1 \ldots dy_n \right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T_i \frac{\partial L}{\partial \beta_j} \, dy_1 \ldots dy_n$$

$$i_j = 1, 2, \ldots, p.$$

where $T = [T_1, T_2, \ldots, T_p]^T$ is a vector function of the elements of Y. The vector T is an unbiased vector estimate for β.

On differentiating both sides of (5.2) with respect to β_i and by condition (i) one obtains

(5.3)
$$0 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial L}{\partial \beta_i} \, dy_1 \ldots dy_n$$

$$0 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{L} \frac{\partial L}{\partial \beta_i} L \, dy_1 \ldots dy_n \right.$$

$$0 = E\left[\frac{\partial \ln L}{\partial \beta_i}\right] \quad i = 1, 2, \ldots, p. \tag{5.4}$$

Let the $p \times 1$ vector S of random variables be defined by

$$S = \left\{\frac{\partial \ln L}{\partial \beta_i}\right\}_p^1 \tag{5.5}$$

with mean vector by (5.4)

$$E(S) = \Phi \tag{5.6}$$

a $p \times p$ covariance matrix $E(SS^T) = \Lambda$, that is

$$\Lambda = \left\{E\left[\frac{\partial \ln L}{\partial \beta_i} \frac{\partial \ln L}{\partial \beta_j}\right]\right\}_p^p \tag{5.7}$$

a positive definite matrix.

Let T be an unbiased vector estimate of β, then

$$\beta_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T_i L \, dy_1 \ldots dy_n \quad i = 1, 2, \ldots, n. \tag{5.8}$$

On differentiating with respect to β_j both sides of (5.8), it follows if condition (ii) holds that

$$\delta_{ij} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T_i \frac{\partial L}{\partial \beta_j} dy_1 \ldots dy_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T_i \frac{\partial \ln L}{\partial \beta_j} L \, dy_1 \ldots dy_n \tag{5.9}$$

$$= E[T_i S_j]$$

where $\delta_{ij} = 1$, if $i = j$ and $\delta_{ij} = 0$, if $i \neq j$.

Let the covariance matrix of T be Σ, that is

$$\Sigma = E[(T-\beta)(T-\beta)^T] \tag{5.10}$$

a positive definite matrix.

Consider the $2p \times 1$ vector v which is built by adjoining the vectors T and S,

$$v = (T^T \vdots S^T)^T \tag{5.11}$$

The covariance matrix R_v of v follows from (5.7), (5.9) and (5.10) and is

$$R_v = \begin{bmatrix} \Sigma & I \\ I & \Lambda \end{bmatrix} \tag{5.12}$$

since $E(TS^T) = I$ is the covariance matrix of the vectors T and S. We note that R_v is positive definite, which implies R_v^{-1} is positive definite. Let R_v^{-1} be partitioned compatibly with (5.11), that is

$$R_v^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where each submatrix A_{ij} are $p \times p$ matrices. The inverse of the principal submatrix [5]

$$A_{11}^{-1} = \Sigma - I \Lambda^{-1} I = \Sigma - \Lambda^{-1}$$

is also positive definite, hence we conclude that

$$\Sigma > \Lambda^{-1} \tag{5.13}$$

the desired matrix lower bound.

5.2 An Application

Let the $n \times 1$ observation vector Y be related to the $p \times 1$ vector of parameters according to the linear model

(5.14) $Y = X\beta + e$

where X is a known $n \times p$ mapping matrix and e is a $n \times 1$ normal random vector, whose mean vector is $E(e) = \phi$, and known $n \times n$ covariance matrix

$$E(ee^T) = R$$

The probability density function of the observation vector Y is

(5.15) $$L = \frac{1}{(2\pi)^{n/2} \mid R \mid^{1/2}} \exp\{-\frac{1}{2}[(Y-X\beta)^T R^{-1} (Y-X\beta)]\}$$

We note from (5.14) that

(5.16) $$E(Y) = X\beta$$

(5.17) $$E(YY^T) = R + X\beta\beta^T X^T.$$

and from (5.15) that

(5.18) $$\frac{\partial \ln L}{\partial \beta} = X^T R^{-1} Y - X^T R^{-1} X\beta$$

It follows that (5.7) and (5.18) that

(5.19) $$\Lambda = E[(X^T R^{-1} Y - X^T R^{-1} X\beta)(X^T R^{-1} Y - X^T R^{-1} X\beta)^T]$$

On expanding the right side of (5.19) and substituting (5.16) and (5.17), the inverse of the matrix bound is found to be

(5.20) $$(X^T R^{-1} X)$$

The desired bound is by (5.13)

$$(X^T R^{-1} X)^{-1}$$

which is clearly the covariance matrix of the vector estimate given by the Gauss-Markov Theorem, that is

$$\hat{\beta} = (X^T R^{-1} X)^{-1} X^T R^{-1} Y.$$

It is immediate that the minimum variance unbiased vector estimate for $C^T \beta$, where C is a vector of constants, is simply $C^T \hat{\beta}$.

5.3 References

[1] Kendall, M. G. and Stuart, S., The Advanced Theory of Statistics, Vol. 2, Griffin and Co. (1961, pp. 9-19.

[2] Hogg, H. V. and Craig, A. T., Introduction to Mathematical Statistics, 2nd ed., Macmillan Co., (1965), pp. 237-242.

[3] Lehman, E. L., Notes on the Theory of Estimation, University of California Press, (1962), pp. 2.3-2.19.

[4] Cramer, H., "A Contribution to the Theory of Statistical Estimation," Skand. Akwarietidskrift, Vol. 29 (1946), p. 85.

[5] Graybill, F. A., An Introduction to Linear Statistical Models, Vol. 1, McGraw-Hill (1961), pp. 8 and 114-117.

[6] Anderson, T. W., "Least Square and Best Unbiased Estimates," Annals of Math. Stat., Vol. 33, No. 1, March 1962, p. 272.

Chapter VI

BEST LINEAR UNBIASED ESTIMATION BY RECURSIVE METHODS WHEN THE OBSERVATIONS ARE CORRELATED

In this chapter a set of recursive formulas will be developed for obtaining the (B.L.U.) estimator for a large class of error correlation models as described in equation (6.4). The advantage of these formulas is that the storage and computational requirements are greatly reduced over the classical solution when the amount of data is much larger than the parameter C. For the case of stationary errors the recursive formulas are only slightly more complicated to implement than the least squares solution for moderate values of C. Thus by using the formulas developed in this chapter, B.L.U. estimators can be used in practice in the later case without undue computational penalties.

6.1 Correlation Model

Let Y be an $n \times 1$ observation vector such that the i^{th} element y_i of Y is an observed scalar at t_i, $i = 1,2,\ldots,n$. Let

$$Y = HX + V \tag{6.1}$$

where H is an $n \times h$ known matrix, X is an $h \times 1$ vector of parameters, and V is an $n \times 1$ error vector such that

$$\begin{aligned} &EV = \phi \\ &E(VV^T) = \rho = (\rho(i,j)) \quad \text{and} \quad \rho(i,j) = E(v_i, v_j), \end{aligned} \tag{6.2}$$

v_i, v_j are the i^{th} and j^{th} elements of V. We note that v_i is not necessarily a stationary process nor are the observations necessarily equally spaced. However, whatever the set of spacing $t^1 = (t_1, t_2, \ldots, t_n)$ of the observations are, the corresponding matrixes H and ρ must be known. The indicated notations of equations merely maps $t_i \rightarrow i$ for convergence of notation.

The problem is to obtain a Best Linear Unbiased Estimate (Markov Estimate) of X, denoted by $\hat{X}$. By definition ρ is a real and symmetric and it will be assumed positive definite. The solution for $\hat{X}$ is well-known [1] and is given by

$$\hat{X} = (H^T \rho^{-1} H)^{-1} H^T \rho^{-1} Y. \tag{6.3}$$

Although the Markov estimate is an optimum estimator, its application in practice is severely limited by the requirements of obtaining ρ^{-1} for large n.

In the material to follow recursive solutions for $\hat{X}$ will be considered. These recursive solutions will considerably reduce the computation and storage requirements as compared to a non-recursive solution.

In the present discussion the correlation matrix is restricted to one of the following cases:

(6.4) a: $\rho(i,j) = 0$ whenever $|i-j| \geq C$

b: $\rho(i,j)$ satisfies a linear difference equation of order C with variable coefficients.

For 6.4a it will be assumed that the covariance matrix ρ is known. For 6.4b one may assume that the coefficients of the difference equation are known and derive the $\rho(i,j)$ satisfying the difference equation or one may be given ρ and derive the coefficients.

The general solution will be presented for the non-stationary process. However if v_i is stationary then the solutions obtained are greatly simplified. The solution for 6.4a will be presented first. It will be shown that the computational difficulties increase quadratically with C for case 6.4a. Thus if condition 6.4a is not met by the actual covariance function, one can take successively larger values of C to obtain a better representation of the process (v_i). One then must weigh the gain in statistical accuracy against the increase in computational complexity.

6.2 The Solution

Since ρ is positive definite, then so is ρ^{-1}. Thus ρ and ρ^{-1} can be uniquely decomposed into

$$\rho^{-1} = S^T S, \quad \rho = S^{-1}(S^T)^{-1} \tag{6.5}$$

where S is a lower triangle matrix with positive diagonal elements. Substituting (6.5) into (6.3) we obtain

$$\hat{X} = (H^T S^T S H)^{-1} H^T S^T S Y. \tag{6.6}$$

Letting

$$H^T S^T = \beta^T \tag{6.7}$$

$$SY = Z,$$

then (6.6) becomes

$$\hat{X} = (\beta^T \beta)^{-1} \beta^T Z. \tag{6.8}$$

Equation (6.8) is known as the least squares estimate of $\hat{X}$ corresponding to the transformed observational vector Z. Thus equation (6.1) becomes under the transformation 6.7

$$Z = \beta X + \omega, \tag{6.9}$$

where $\omega = SV$. The covariance of ω is given by

$$E(\omega\omega^T) = E(SVV^T S^T) = S\rho S^T$$

$$= (S\, S^{-1})(S^{T^{-1}} S^T)$$

or

$$E(\omega\omega^T) = I. \tag{6.10}$$

Under the transformation 6.7 the problem has been reduced to determining the transformation S from a given covariance matrix ρ and manipulating the equations to minimize the storage requirements of components of S, H, and Y.

6.3 Determination of the Transformation S

One may determine the matrix S satisfying 6.5 in the following way. Consider

$$\omega = SV.$$

Let $S^{-1} = A\Sigma$, where A is lower triangle and Σ diagonal, which implies

$$V = A\Sigma\omega, \tag{6.11}$$

where $EV = \phi$, $E\omega = \phi$, $E(VV^T) = \rho = (\rho(i,j))$, and $E(\omega\omega^T) = I$. Equation 6.11 can be written as

$$v_m = \sum_{j=1}^{m} a_{mj}\sigma_j\omega_j, \quad m = 1,2,\ldots,n. \tag{6.12}$$

Let $A_{jj} = 1$, $j = 1,2,\ldots,n$. Thus it follows

$$E(v_m^2) = \rho(m,m) = \sum_{j=1}^{m-1} A_{mj}^2\sigma_j^2 + \sigma_m^2$$

which implies

$$\sigma_m^2 = \rho(m,m) - \sum_{j=1}^{m-1} A_{mj}^2\sigma_j^2, \quad m = 1,2,\ldots,n. \tag{6.13}$$

Now if one considers

$$E(v_m v_1) = E[(\sum_{j=1}^{m} a_{mj}\sigma_j\omega_j)(a_{11}\sigma_1\omega_1)]$$

$$= a_{m1}a_{11}\sigma_1^2 = a_{m1}\rho(1,1)$$

then

$$A_{m1} = \frac{\rho(m,1)}{\rho(1,1)}, \quad m = 1,2,\ldots,n. \tag{6.14}$$

Also

$$E(v_m v_i) = E[(\sum_{j=1}^{m} A_{mj}\sigma_j\omega_j)(\sum_{t=1}^{i} A_{it}\sigma_t\omega_t)]$$

$$\rho(m,i) = A_{m1}A_{i1}\sigma_1^2 + \ldots + A_{mi}A_{ii}\sigma_i^2$$

which implies

$$A_{mi} = \frac{(m,i) - \sum_{j=1}^{i-1} A_{mj}A_{ij}\sigma_i^2}{\sigma_i^2}, \quad m \geq i, \; L = 1,\ldots,m. \tag{6.15}$$

Thus the above formulas allows one to calculate the coefficient A_{mi} recursively given $\rho(i,j)$. If (6.4a) holds a simplification of the above formula results. Assume (6.4a) holds, then

$$m - i \geq C$$

implies from the above formulas that $A_{mi} = 0$. For suppose t is such that $m - t \geq C$ and $m - t - 1 < C$, then formula (6.14) implies $A_{m1} = 0$ if $m - 1 \geq C$ and from formula (6.15) $A_{mi} = 0$ if $m - i \geq C$. Hence $A_{mi} = 0$ when $m - i \geq C$. Thus (6.12) becomes

$$v_m = \sum_{j=\max(1,1+m-C)}^{m} A_{mj}\sigma_j\omega_j \quad , \quad m = 1,2,3,\ldots, \tag{6.16}$$

and if $n - m \geq C$, then $m < 1 + m - C$ implies

$$E(v_n,v_m) = E[(\sum_{j=\max(1,1+n-C)}^{m} A_{nj}\sigma_j\omega_j)\quad(\sum_{m=\max(1,1+m-C)}^{m} A_{mj}\alpha_j\omega_j)\quad]$$

$$= 0. \tag{6.17}$$

Thus A is a lower triangle matrix with at most C, non-zero diagonals, i.e.

$$A_{mj} = 0, \quad (m-j) \geq C \tag{6.18}$$

$$(m-j) < 0.$$

6.4 Computation of the m^{th} Row of A

Let $m \geq i$, $i = m, m - 1, m - 2, \ldots, m - C + 1$; $m = 2,3,\ldots$.

Then by applying (6.18) to (6.15) we obtain

$$A_{mi} = \frac{\rho(m,i) - \sum_{j=\max(1,m-C+1}^{i-1} A_{mj}A_{ij}\sigma_j^2}{\sigma_i^2} . \tag{6.19}$$

Since $A_{mi} = 0$ when $m - i \geq C$ then the computation of the m^{th} row, $A_{(m)}$, of A requires the storage of at most the last C rows of A that is $A_{(m)}, A_{(m-1)}, \ldots, A_{(m-C+1)}$, and the sequence $\{\sigma_j^2\}$, $j = m, m - 1, m - a, \ldots, m - C + 1$.

Since each row $A(u)$, $u \geq C$, contains (C-1) non-zero coefficients A_{ui}, i.e. $A_{uu}, A_{u,u-1}, \cdots, A_{u,u-C+1}$ are possibly different from zero but $A_{u,u-C}, A_{u,u-C-1}, \cdots, A_{u,1}$ are all zero, then the storage is of the order of $(C-1)^2$.

6.5 Computation of (σ_j^2)

Equation (6.13) becomes

$$\sigma_m^2 = \rho(m,m) - \sum_{j=\max(1,m-C+1)}^{m} A_{mj}^2 \sigma_j^2 \quad . \tag{6.20}$$

Now is $i = m - C + 1$ in (6.19) then we must require the storage of the sequence $\{\sigma_u^2\}$, $u = m - 2C + 2, m - 2C + 1, \ldots, m - C + 1$. Thus one sees that the storage requirements for the m^{th} row of the matrix A increases as C^2 and is not a function of m. Hence for large m this represents a substantial storage savings. These properties follows as a consequence of property (6.4a).

6.6 Inversion of A Matrix

The discussion so far has considered recursive methods of evaluating the elements A_{mi} of A. However from the equation $\omega = SV$ one sees that the desired transformation S is in terms of A^{-1}. Therefore we will now consider properties of the elements b_{kr} of A^{-1}.

Let

$$B = A^{-1} = \{b_{kr}\} \qquad \begin{matrix} r = 1,2,3,\ldots,n \\ k = 1,2,3,\ldots,n \end{matrix}$$

Since $A = \{A_{ij}\}$ is a lower triangle matrix it follows that

(a) A^{-1} is lower triangle matrix

(b) $A_{jj} = 1$ implies $b_{kk} = 1$,

and finally let $k > r$ then

$$b_{kr} = -\sum_{j=r}^{k-1} b_{jr}A_{kj} \qquad \begin{matrix} r = 1,2,\ldots,k-1 \\ k = 2,3,\ldots,n \end{matrix} \quad . \tag{6.21}$$

To see formula (6.21) multiply the k^{th} row of A times the r^{th} column of B. Thus we get

$$A_{kr}b_{rr} + A_{kr+1}b_{r+1,r} + \cdots + A_{k_1k-1},b_{r-1,r} + A_{kk}b_{,,r} = 0$$

or

$$b_{kr} = \sum_{j=r}^{b-1} b_{jr}A_{kj}.$$

$$b_{kk} = 1, \quad k = 1,2,\ldots,n \tag{6.22}$$

$$b_{kr} = 0, \quad k < r. \tag{6.23}$$

Using formula (6.18), we note that formula (6.21) becomes

$$b_{kr} = -\sum_{j=k-C+1}^{k-1} b_{jr}A_{kj}. \tag{6.24}$$

Equation (6.24) leads to a very useful relationship between the observed sequence $\{y_1\}$ and the transformed sequence $Z = SY$. Thus one has $Z = \Sigma^{-1}A^{-1}Y$ which implies

$$Z_k = \frac{1}{\sigma_k} \sum_{r=1}^{k} b_{kr} y_r, \qquad k = 1,2,\ldots,n. \tag{6.25}$$

$$= \frac{1}{\sigma_k}[y_k + \sum_{r=1}^{k-1} b_{kr} y_r]. \tag{6.26}$$

Substituting (6.24) into (6.26) one obtains

$$Z_k = \frac{1}{\sigma_k}[y_k - \sum_{j=k-C+1}^{k-1} \sum_{r=1}^{k-1} b_{jr} A_{kj} y_r] \tag{6.27}$$

Fix $j = k-1, k-2, \ldots, k-C+1$ and noting (6.25) for $k > C$, then

$$Z_k = \frac{1}{\sigma_k}[y_k - \sum_{r=1}^{k-1} b_{k-1,r} A_{k,k-1} y_r - \sum_{r=1}^{k-1} b_{k-2,r} A_{k,k-2} y_r - \cdots$$

$$- \sum_{r=1}^{k-1} b_{k-C+1,r} A_{k,k-C+1} y_r]$$

$$= \frac{1}{\sigma_k}[y_k - A_{k,k-1}\sigma_{k-1} Z_{k-1} - A_{k,k-1}\sigma_{k-2} Z_{k-2}$$

$$- A_{k,k-C+1}\sigma_{k-C+1} Z_{k-C+1}]$$

$$= \frac{1}{\sigma_k}[y_k - \sum_{u=1}^{C-1} A_{k,k-u}\sigma_{k-u} Z_{k-u}] \tag{6.28}$$

For $k \leq C$, take $y_v = 0$ if $v < 0$.

Equation (6.28) shows that to generate the k^{th} transformed variable Z_k one requires storage of the previous C values of Z and the coefficients of the k^{th} row of A as well as the present and C previous values of σ_k. Thus as n gets very large and one wishes to compute Z_{n+1}, the transformed data need be stored over a span of C observations

rather than n, since Z_n satisfies a linear difference equation with non-constant coefficients given by (6.28).

6.7 Recursive Relationships for the Parameter Estimates

Let $\hat{X}^{(n)}$ be the Markov estimate based on an n^{th} dimensional observational vector, and define $H^{(n)}$ correspondingly. Let us assume that an estimate $\hat{X}^{(n)}$ has been obtained and that m additional observations have been made, and it is desired to obtain a Markov estimate of X^T based on the $(m+m)$ observations recursively in terms of $\hat{X}^{(n)}$.

The first n^2 terms $\rho(i,j)$, $i,j = 1,2,\ldots,n$ of the covariance matrix do not change as ρ goes from an $n \times n$ to an $(n+m \times (n+m)$ matrix. Similarly the first n rows of matrix H of equation (6.1) are unaltered by the addition of m rows to H. Thus the first n rows of matrices A and A^{-1}, and Σ are also unaltered as these matrices change from $n \times n$ to $(n+m) \times (n+m)$ matrices by the addition of m more observations. The net effect is that the first n rows of the Z vector and the first n rows of the β matrix are unaltered. Thus let $\beta^{(u)}$ be the β matrix based on the first u observations and similarly for $Z^{(u)}$. We may write (6.6) as

$$\hat{X}^{(m+n)} = (\beta^{(m+n)^T} \beta^{(m+n)})^{-1} \beta^{(m+n)^T} Z. \tag{6.29}$$

We may partition matrices β and Z as follows,

$$(6.30) \qquad \beta^{(m+n)T} = \underset{h \times n}{\left[\beta^{(n)T}\right.} \;\Big|\; \underset{h \times m}{\left.\delta\beta^{T}\right]}$$

$$(6.31) \qquad Z^{(m+n)} = \underset{m \times 1}{\begin{bmatrix} Z^{(n)} \\ \hline \delta(Z) \end{bmatrix}} \quad n \times 1$$

Thus (6.29) becomes

$$(6.32) \qquad \hat{X}^{(m+n)} = [\beta^{(n)T}\beta^{(n)} + \delta\beta^{T}\delta\beta]^{-1}[\beta^{(n)T}Z^{(n)} + \delta\beta^{T}\delta Z]$$

$$= (I + \Delta)^{-1}(\beta^{(n)T}\beta^{(n)})^{-1}\,[\beta^{(n)T}Z^{(n)} + \delta\beta^{T}\delta Z].$$

$$(6.33) \qquad \hat{X}^{(m+n)} = (I + \Delta)^{-1}[\hat{X}^{(n)} + (\beta^{(n)T}\beta^{(n)})^{-1}\delta\beta^{T}\delta Z]$$

where

$$(6.34) \qquad \Delta = (\beta^{(n)T}\beta^{(n)})^{-1}\delta\beta^{T}\delta\beta.$$

We note that Δ is at least positive semi-definite and has eigen values greater than or equal zero. Let λ be the maximum eigen value of Δ. Then [15] if $\lambda < 1$, $(I + \Delta)^{-1}$ permits an expansion of the form

$$(6.35) \qquad (I + \Delta)^{-1} = I - \Delta + \Delta^{2} - \Delta^{3} + \ldots .$$

If $\lambda << 1$, then

$$(I + \Delta)^{-1} \simeq I - \Delta. \tag{6.36}$$

In many applications if $n >> m$, $(\beta^{(m)^T}\beta^{(n)})$ is ill conditioned and therefore (6.35) is a very useful approximation for obtaining

$$(\beta^{(n+m)^T}\beta^{(n+m)})^{-1} \simeq (I - \Delta)(\beta^{(n)^T}, \beta^{(n)})^{-1}, \tag{6.37}$$

and for direct substitution into (6.33).

6.8 Recursive Relationship for Elements of $\delta\beta$

From (6.7) and $S^{-1} = A\Sigma$ we may write

$$\beta = SH = \Sigma^{-1}A^{-1}H = (\beta_{u,v}). \tag{6.37}$$

The elements β_{uv} of β are given by

$$\beta_{uv} = \frac{1}{\sigma_u}\sum_{j=1}^{u} b_{uj}h_{jv} \tag{6.38}$$

$$= \frac{1}{\sigma_u}[h_{uv} + \sum_{j=1}^{u-1} b_{uj}h_{jv}]. \tag{6.39}$$

Substituting (6.24) into (6.38) we obtain

$$\beta_{uv} = \frac{1}{\sigma_u}[h_{uv} + \sum_{i=u-C+1}^{u-1} A_{ui} \sum_{j=1}^{u-1} b_{ij}h_{jv}].$$

Fix $i = u-1, u-2, \ldots, u-C+1$ and noting (6.38) for $u \geq C$, then

$$\beta_{uv} = \frac{1}{\sigma_u}[h_{uv} - A_{u,u-1}\sigma_{u-1}b_{u-1,\,v} - A_{u,u-2}\sigma_{u-2}b_{u-2,\,v}$$

$$- \cdots A_{u,u-C+1}\sigma_{u-C+1}b_{u-C+1}]$$

$$= \frac{1}{\sigma_u}[h_{uv} - \sum_{j=1}^{C-1} A_{u,u-j}\beta_{u-j,v}\sigma_{u-j}]. \tag{6.40}$$

That is for each column of β, the elements in the v^{th} column satisfy the same difference equation as do the Z variables of (6.28). To obtain the elements of $\delta\beta$ we note

$$\delta\beta = (\beta_{u,v}) \qquad \begin{matrix} u = m+1,\ldots,n+m \\ v = 1,2,\ldots,h \end{matrix}$$

It is therefore only necessary to store additionally the C previous rows of the matrix $\beta^{(n)}$.

6.9 Stationary Case

Let us assume that condition (6.4a) holds and in addition

$$\rho(i,j) = \rho(|i-j|).$$

Then from (6.19) with $m \geq i \geq C$

$$A_{mi}\sigma_i^{\,2} = \rho(m,i) - \sum_{j=m-C+1}^{i-1} A_{mj}A_{ij}\sigma_j^{\,2}$$

which implies

$$\rho(m,i) = \sum_{j=m-C+1}^{i} A_{mj}A_{ij}\sigma_j^{\,2} \tag{6.41}$$

$$= \rho(m-i)$$

$$\rho(m+1,i+1) = \rho(m,i) = \sum_{j=m-C+2}^{i+1} A_{m+1,j}A_{i+1,j}\sigma_j^2 . \tag{6.42}$$

Equating (6.41) and (6.42) we obtain that ω sufficient condition for (6.41) and (6.42) to be satisfied in that

$$A_{mj}\sigma_j = A_{m+1,j+1}\sigma_{j+1} . \tag{6.43}$$

Formula (6.43) follows by letting $t = j-1$ in formula (6.42). Since formula (6.43) holds for $j = m$, then

$$\sigma_j = \sigma_{j+1} = \sigma, \; j \geq m \geq C \tag{6.44}$$

$$A_{mj} = A_{m+1,j+1} \qquad j = m, m-1, \ldots, m\_C+1 .$$

Thus each diagonal of matrix A has constant value from the C^{th} row. The elements of δZ and $\delta\beta$ are easily obtained as follows. Let

$$A_{uv} = A_{uv,u\ C,v=u,u-1,\ldots,u-C+1} . \tag{6.45}$$

Then from (6.28)

$$Z_k = \frac{1}{\sigma}[y_k - \sum_{u=1}^{C-1} A_{k,k-u}\ \sigma Z_{k-u}] \tag{6.45}$$

$$= \frac{y_k}{\sigma} - \sum_{u=1}^{C-1} A_{k,k-u}Z_{k-u} , \quad k \geq 2C$$

and from (6.40)

$$\beta_{u,v} = \frac{h_{uv}}{\sigma} - \sum_{j=1}^{C-1} A_{u,u-j}\beta_{u-j,v} , \; u \geq 2C . \tag{6.46}$$

The computations of δZ, $\delta\beta$, δ, and $X^{(n+n)}$ proceeds similiar to the above. 6.8 $\rho(i,j)$ satisfied a difference equation.

6.10 $\rho(i,j)$ Satisfied a Difference Equation

We will now develop recursive estimates of $\hat{X}^{(n+m)}$ based on condition (6.4b): First the general case of non-stationary (v_i) will be considered. The simplification of the solution when (v_i) is stationary will be shown. An approximate solution will be demonstrated for a class of processes which are asymptotically stationary. Finally the nature of the covariance matrix will be discussed when the time varying coefficients of the difference equations are taken as constant (model of reference 12). The approach will be to obtain a partitioning of ρ^{-1} into a lower triangular matrix α such that $\rho^{-1} = \alpha^T\alpha$. The procedure then for obtaining $\hat{X}^{(n+m)}$ will be as previously shown. The problem will be considered from two points of view. In the first case it is assumed that the coefficients of the difference equation are given and it is required to generate the elements of ρ. In the second case it will be assumed that one is given the elements of ρ and is required to generate α. In the latter case it will be required to insert matrices of order $\min(u,c)$ for row u of α. However, when (v_i) is stationary or the diagonals of α are constant then all rows of α are equal for $u \geq c$ and only the inversion of one matrix of order c is required to obtain all rows of α for $u \geq c$.

Let the random sequence $\{v_i\}$ satisfy

$$\omega_u = \sum_{j=\max(1,u-C+1}^{u} \alpha_{ju} v_j \qquad u = 1,2,\ldots,n \tag{6.47}$$

with $\alpha_{uu} > 0$ for each α_u and α_{uj} real.

That is $\alpha = (\alpha_{ij})$ is a lower triangle matrix (zeros above the main diagonal) with at most C non zero diagonals. Thus (6.47) can be written as

$$\omega = \alpha v. \tag{6.48}$$

Now it follows that $E(\omega) = \phi$ and $E(\omega\omega^T) = I$. Hence we can write

$$E(\omega_u^2) = \sum_{j,k=\max(1,u-c+1)}^{u} \alpha_{uj}\alpha_{uk}\rho(j,k) \tag{6.49}$$

It follows that

$$V = \alpha^{-1}\omega. \tag{6.50}$$

Multiplying each side of (6.47) by v_k, $u \geq k$ then using

$v_j = \sum_{i=1}^{j} d_{ji}\omega_i$ where $\{d_{ji}\} = \alpha^{-1}$ one finds

$$E(v_k\omega_u) = \sum_{j=\max(1,u-c+1)}^{u} d_{uj}\rho(j,k). \quad \text{But}$$

$$E(v_k,\omega_u) = E(\sum_{i=1}^{k} d_{ki}\omega_i\omega_u) = \sum_{i=1}^{k} d_{ki}\delta_{iu}$$

$$= d_{kk}\delta_{ku}$$

which implies

$$d_{uu} = \sum_{j=\max(1,u-c+1)}^{u} \alpha_{uj}\rho(j,u) \quad . \tag{6.51}$$

Thus we note (6.51) shows that the covariance function $\rho(j,u)$ satisfies a difference equation with time varying coefficients α_{uj} and forcing function d_{uu}. To evaluate d_{uu} note that

$$E(VV^T) = \rho = \alpha^{-1}(\alpha^{-1})^T \tag{6.52}$$

or

$$\sum_{j=\max(v,u-c+1)}^{u} \alpha_{uj} d_{jv} = \delta_{uv}, \quad v \leq u. \tag{6.53}$$

Letting $u = v$ then

$$\alpha_{vv} d_{vv} = 1 \tag{6.54}$$

which implies

$$d_{vv} = \alpha_{vv}^{-1}. \tag{6.55}$$

Similarly for $u = 1,2,3,\ldots$, $k = 1,2,\ldots,u-1$

$$\sum_{j=\max(u-k,u-c+1)}^{u} \alpha_{uj} d_{j,j-k} = 0 \tag{6.56}$$

or

$$d_{u,u-k} = -d_{uu} \sum_{j=\max(u-k,u-c+1)}^{u-1} \alpha_{uj} d_{j,u-k} \,. \tag{6.57}$$

6.11 Generation of Covariance Matrix Given α.

Given the coefficients α_{uj} of α one can generate the elements of ρ as follows. Since

$$d_{kk}\delta_{ku} = \sum_{j=\max(1,u-c+1)}^{u} \alpha_{ui}\rho(j,k) \quad ,$$

then for $k = 1$

$$\frac{1}{\alpha_{11}}\delta_{1u} = \sum_{j=\max(1,u-c+1)}^{u-1} \alpha_{uj}\rho(j,1) + \alpha_{uu}\rho(u,1)$$

or

$$\rho(u,1) = \frac{1}{\alpha_{uu}} - \sum_{j=\max(1,u-c+1)}^{u-1} \alpha_{uj}\rho(j,1) + \frac{\delta_{1u}}{\alpha_{11}} \quad . \tag{6.58}$$

Therefore

$$\rho(1,1) = \frac{1}{\alpha_{11}^{2}} \quad . \tag{6.59}$$

One may then compute $\rho(1,u)$, $u = 2,3,\ldots,n$ recursively from (6.58). Also the $u = k$

$$\frac{1}{\alpha_{kk}} = \sum_{j=\max(1,k-c+1)}^{k} \alpha_{kj}\rho(j,k) \quad . \tag{6.60}$$

For $k = 2$

$$\rho(2,2) = \alpha_{22}^{-1}\,(-\alpha_{21}\,(2,1) + \alpha_{22}^{-1}). \tag{6.61}$$

Since $\rho(2,1)$ is obtained from (6.58), we may solve for $\rho(2,2)$. We may now generate for $u = 3,4,\ldots$.

$$\rho(2,u) = -\frac{1}{\alpha_{22}} \sum_{j=\max(1,u-c+1)}^{u-1} \alpha_{uj}\rho(2,j) \quad . \tag{6.62}$$

By a similar procedure we may solve for each $\rho(u,u)$ from (6.60) and obtain

$$\rho(v,u) = -\frac{1}{\alpha_{uu}} \sum_{j=\max(1,u-c+1)}^{u-1} \alpha_{uj}\rho(j,v) \quad , \quad u = v+1,v+2,\ldots,n. \tag{6.63}$$

6.12 Generation of α Given the Covariance Matrix ρ

The problem of more practical importance is the generation of the matrix α from a given ρ. In general given a positive definite covariance matrix ρ, the inverse ρ^{-1} is also positive definite and there exists a unique decomposition of ρ^{-1} as given by

$$\rho^{-1} = \alpha^{T}\alpha \tag{6.64}$$

into the product of a triangular matrix and its transpose, where α has positive diagonal elements. This requires in general that the summation of (6.47) be from $j = 1$ to u. The effect of the transformation (6.47) is to require that $\alpha_{ij} = 0$ whenever $(i-j) \geq c$, so that the difference equation which is satisfied by $\rho(i,j)$ is of maximum order C. If this condition is not actually met then (6.4b) can be considered as a C^{th} order approximation to the true difference equation whose order may grow indefinitely large with n. This approximation would seem reasonable to use in those cases where the

magnitude of the difference equation coefficients α_{ij} drops rapidly as $i-j = (c,c+1,\ldots)$ whereas the correlation coefficients $\rho(i,j)$ may not decrease rapidly as $|i-j| = (c,c+1,\ldots)$. If on the other hand $\rho(i,j) \to 0$ rapidly as $|i-j| = (c,c+1,\ldots)$ then the methods based on assumption (6.4a) would be more useful.

6.13 Computation of α_{uj}, $u \leq C$

Given the $\rho(u,m)$ one can obtain the α_{ij} as follows: From

$$E(v_k\omega_u) = d_{kk}\delta_{ku} = \sum_{j=\max(1,u-c+1)}^{u} \alpha_{uj}\rho(j,k) \tag{6.65}$$

we may write for $k = 1$, $u = 1$

$$\alpha_{11}^{-1} = \alpha_{11}\rho(1,1)$$

or

$$\alpha_{11} = \rho(1,1)^{-1/2}.$$

Letting $k = 1$, $u = 2$

$$0 = \sum_{j=1}^{2} \alpha_{2j}\rho(j,1)$$

or

$$\alpha_{21}\rho(1,1) + \alpha_{22}\rho(2,1) = 0.$$

Letting $k = 2$, $u = 2$, then

$$\alpha_{22}^{-1} = \sum_{j=1}^{2} \alpha_{2j}\rho(j,2)$$

or

$$\alpha_{21}(1,2) + \alpha_{22}\rho(2,2) = \alpha_{22}^{-1}.$$

Now by multiplying through by α_{22} one obtains the system

(6.66)
$$\alpha_{21}\alpha_{22}\rho(1,1) + \alpha_{22}^{2}\rho(2,1) = 0$$

$$\alpha_{21}\alpha_{22}\rho(1,2) + \alpha_{22}^{2}\rho(2,2) = 1.$$

Let

(6.67) $\beta_{ij} = \alpha_{ij}\alpha_{ii}$, then

(6.68)
$$\begin{bmatrix} \rho(1,1) & \rho(2,1) \\ \rho(1,2) & \rho(2,2) \end{bmatrix} \begin{bmatrix} \beta_{21} \\ \beta_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Let $\rho^{(2)} = \begin{bmatrix} \rho(1,1) & \rho(2,1) \\ \rho(1,2) & \rho(2,2) \end{bmatrix}$, then

(6.69)
$$\beta_{21} = \frac{-\rho(2,1)}{\det \rho^{(2)}}$$

$$\beta_{22} = \frac{\rho(1,1)}{\det \rho^{(2)}} = \alpha_{22}^{2}$$

where $\det A$ is the determinant of A.

The construction continues for $\beta_{3.}^{T} = (\beta_{31},\beta_{32},\beta_{33})$

$$\beta_{3.} = \rho^{(3)^{-1}} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{6.70}$$

up to $\beta_{C.}^{T}$ where each of the matrices $\rho(u)$ is given by

$$\rho^{(u)} = (\rho(i,j)),\ i,j = 1,2,\ldots,u, \text{ for } u \leq C. \tag{6.71}$$

Therefore

$$\beta_{u.} = \rho^{(u)^{-1}} \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix} . \tag{6.72}$$

Let $R_{ij}^{(u)}$ be the cofactor of the element $\rho(j,j)$ of $\rho^{(u)}$. Then

$$\beta_{uj} = \frac{R_{uj}^{(u)}}{\det \rho^{(u)}} , \tag{6.73}$$

and since

$$\beta_{uu} = \alpha_{uu}^{2} ,$$

we may determine

$$\alpha_{uu} = \beta_{uu} \tag{6.74}$$

and then α_{uj}, $j=1,2,\ldots,u-1$ from (6.67).

6.14 Computation of α_{vj}, $v \geq C$

Define

(6.75) $$\rho^{(v)} = (\rho(i,j)),\ i,j = v-c+1,\ v-c+2,\ \ldots,\ v.$$

When $v \geq c$, (6.65) may be written

$$\delta_{kv} = \frac{1}{d_{vv}} \sum_{j=v-c+1}^{v} \alpha_{vj}\rho(j,k)$$

$$= \sum_{j=v-c+1}^{v} \alpha_{vv}\alpha_{vj}\rho(j,k)$$

(6.76) $$\alpha_{kv} = \sum_{j=v-c+1}^{v} \beta_{vj}\rho(j,k),\quad v \geq k.$$

Let $k = v$

$$1 = \sum_{j=v-c+1}^{v} \beta_{vj}\rho(j,v),$$

$k = v-1$

$$0 = \sum_{j=v-c+1}^{v} \beta_{vj}\rho(j,v-1),$$

$\vdots$

$k = v-c+1$

$$0 = \sum_{j=v-c+1}^{v} \beta_{vj}\rho(j,v-c+1)$$

or

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \rho(v-c+1,v-c+1) & \dots & \rho(v,v-c+1) \\ & & \\ & & \\ \rho(v-c+1,v) & \dots & \rho(v,v) \end{bmatrix} \begin{bmatrix} \beta_{v,v-c+1} \\ \vdots \\ \beta_{vv} \end{bmatrix}$$

$c \times 1 \qquad\qquad c \times c \qquad\qquad c \times 1$

which implies

$$\beta_{v.} = \rho^{(v)^{-1}} \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix} ; \tag{6.77}$$

where $\rho^{(v)}$ is a $c \times c$ matrix whose matrix inverse exists since ρ is positive definite. The components β_{vj}, may now be obtained by using equation (6.73) and (6.74) which hold for $u \geq C$. Thus we see that to obtain the u^{th} row of α for $u \geq C$ it is necessary to invert a $c \times c$ matrix. If C is very large this will make the recursive solution of doubtful value. It will be seen however that the method simplifies considerably if (v_i) is stationary.

6.15 Stationary (v_i)

Let us assume that the sequence (v_i) is a stationary so that

$$\rho(i,j) = \rho(|i-j|). \tag{6.78}$$

Then

$$\rho^{(v)} = \rho^{(v+1)} \quad , \quad v = c, c+1, \ldots$$

and therefore

(6.79) $$\beta_{v.} = \beta_{v+1},$$

and $\alpha_{v,j} = \alpha_{v+1,j+1}$, $j = v, v-1, \ldots$.

Note that α_{vj} is not necessarily equal to $\alpha_{v+1,j+1}$ if $v < C$.

Thus in the stationary case all the diagonal terms are constant for all rows equal or greater than the C^{th} row. It is now only necessary to invert the matrix $\rho^{(c)}$ to obtain all subsequent values of α_{uj}, $u \geq C$. This property is very useful and can be helpful in obtaining a recursive relationship for the α_{uj} for process which though not strictly stationary, approach a stationary condition asymptotically.

6.16 Recursive Parameter Estimates

It is desired to obtain recursive estimates of the Markov estimator $\hat{X}$. The method is the same as previously shown with slight modifications. Let

(6.80) $$\varepsilon = \alpha H$$

(6.81) $$Z = \alpha y.$$

Then from $\rho^{-1} = \alpha^T\alpha$ and (6.3)

$$\hat{X} = (\varepsilon^T\varepsilon)^{-1}\ \varepsilon^T Z. \tag{6.82}$$

Again the first n rows of the Z vector are unaltered and the first n rows of the ε matrix are unaltered as one goes from n to $n + m$ observations. Thus equation (6.30) and (6.37) hold with the substitutions β to ε and Z to Z and Δ to δ. It now remains to determine the elements of $\delta\varepsilon$ and δZ and show that they require storage of at most the last C rows of $H^{(n)}$ and $Y^{(n)}$. Let

$$\varepsilon = (\varepsilon_{ij}) \tag{6.83}$$

then

$$\varepsilon_{ij} = \sum_{t=1}^{n+m} \alpha_{it} H_{t,j} \qquad \begin{matrix} i = 1,2,\ldots,n+m \\ j = 1,2,\ldots,h \end{matrix} \tag{6.84}$$

But

$$\alpha_{i,t} = 0 \quad \text{if} \quad 0 < i-t \leq C.$$

Therefore

$$\varepsilon_{ij} = \sum_{t=\max(i,i-c+1)}^{i} \alpha_{it} H_{t,j} \tag{6.85}$$

and

$$\delta\varepsilon = (\varepsilon_{ij}) \qquad \begin{matrix} i = n+1,\ldots,n+m \\ j = 1,2,\ldots,h. \end{matrix} \tag{6.86}$$

Let $n \geq C$ then

$$\varepsilon_{ij} = \sum_{t=i-c+1}^{i} \alpha_{it} H_{t,j} \qquad \begin{matrix} i = n,n+1,\ldots,n+m \\ j = 1,2,\ldots,h. \end{matrix} \tag{6.87}$$

Therefore to compute the elements of $\delta\varepsilon$ one must store the last C rows of $H^{(n)}$.

Similarly for $i \geq C$

$$Z_i = \sum_{t=i-c+1}^{i} \alpha_{it} y_t \qquad i = c,c+1,\ldots \tag{6.88}$$

requires the storage of at most the last C values of y_n. When (v_i) is stationary then one may substitute $\alpha_{i,t} = \alpha_{i-t,i=c,c+1,\ldots}$ in equation (6.87) and (6.88).

6.17 References

[1] Gauss, Carl Frederick, "Theoria Combinationis observationum Erroribus Minimus Obnoxiac" French Translation "Methodedes Moindres Carres," Memories Sur la Combination des Observation; published in Paris by Bertrand, 1855, Gottingen.

[2] Aitken, A. C., "On Least Squares and Linear Combinations of Observations," Proceedings Royal Society of Edinburgh, 1935, Vol. 1, pg. 42.

[3] Grenander, U., Rosenblatt, M., "Statistical Analysis of Stationary Time Series," John Wiley and Sons, 1957.

[4] Weiner, Norbert, "Extropolation Interpolation and Smoothing of Stationary Time Series," John Wiley and Sons, Inc., New York, 1950.

[5] Kolmogoroff, A., "Interpolation and Extropolation Von Stationaren Zufalligen Folgen," Bulletin de l'academie des sciences de U.R.S.S., Sr. Math. 5, pp. 3-14, 1941.

[6] Blum, M., "Fixed Memory Least Squares Filters Using Recursive Methods," IRE Transactions of the Professional Group on Information Theory, Vol. IT-3, #3, Sept., 1957.

[7] Blum, M., "On Exponential Filters," Journal of the Association for Computing Machinery, Vol. 6, No. 2, pp. 283-304, April, 1959.

[8] Swerling, P., "First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations," Journal of Astronautical Sciences, Vol. 6, No. 3, pp. 46-52, Autum, 1959.

[9] Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Journal of Basic Engineering, Trans. Am. Soc. Mech. Engrs., 820, pp. 34-45, March, 1960.

[10] Battin, R. H., "A Statistical Optimization Navigation Procedure for Space Flight," A.R.S. Journal 32, pp.1681-1696, No., 1962.

[11] Claus, A. J., Blackman, R. B., Halline, E. G., and Ridgeway, W. C., "Orbit Determination and Prediction and Computer Programs," Bell System Technical Journal, (Special Issue on Telstar I), July, 1963.

[12] Blum, M., "A Stagewise Parameter Estimation Procedure for Correlated Data," Numerische Mathematick 3, 1961, pp. 202-208.

[13] Marcus, M., "Basic Theorems in Matrix Theory," National Bureau of Standards Applied Mathematics Series 15, January 22, 1960.

[14] Forsythe, G. E., "Theory of Selected Methods of Finite Matrix Inversion and Decomposition," U. S. National Bureau of Standards, INA-52-5, August 13, 1951.

[15] Freidman, B., "Principles and Techniques of Applies Mathematics," John Wiley and Sons, Inc., New York, 1957.

[16] Blum, M.,"Best Linear Unbiased Estimation by Recursive Methods," SIAM Journ on Applied Math, Vol. 14, No. 1, 1966, pp. 167-168.

Chapter VII

ON SELECTING SAMPLE POINTS

Given the linear model $Y = X\beta + e$ where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

and

$$e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

with $Ee = \emptyset$

and Case I: $Eee^t = I\sigma^2$

Case II: $Eee^t = V\sigma^2$

We wish to look at a procedure to choose the x_i's, $i = 1, \ldots, n$, so that the $\text{cov}(\hat{\beta}) = (X^tV^{-1}X)^{-1}\sigma^2$ is a minimum.

If we allow the x_i to be chosen from an unlimited range, then we can make the $\text{cov}(\hat{\beta})$ arbitrarily small by choosing the X matrix to be aX, so that

$$\text{cov}(\hat{\beta}) = a^{-1}(X^tV^{-1}X)^{-1}\sigma^2 \text{ which approaches } 0 \text{ as } a \to \infty.$$

Thus we will limit the range of the x_i to small regions in which we may be interested.

Then we can make a restriction of the form:

$$(7.1) \qquad X_i^t V^{-1} X_i = C_i^2 \text{ (given)}, \quad i = 1,2.$$

where X_i is the i^{th} column of X.

THEOREM 7.1 Let X be a design matrix and the $\hat{\beta}_i$ be the least squares estimator of β_i. Then, under the restrictions in (1) on X,

(a) $V(\beta_i) \geq \frac{1}{C_i^2}$

(b) the minimum is attained when $(X_i^t V^{-1} X_j) = 0$, $i \neq j$.

Proof: Let $i = 1$. The matrix $X^T V^{-1} X$ can be written as follows

$$(7.2) \qquad X^T V^{-1} X = \begin{bmatrix} X_1^T V^{-1} X_1 & X_1^T V^{-1} X_2 \\ X_2^T V^{-1} X_1 & X_2^T V^{-1} X_2 \end{bmatrix},$$

where $X = (X_1, X_2)$.

Since $X^T V^{-1} X$ is positive definite, then

$$(X^T V^{-1} X)^{-1} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

where

$$R_{11} = [(X_1^T V^{-1} X_1) - (X_1^T V^{-1} X_2)(X_2^T V^{-1} X_2)^{-1} X_2^T V^{-1} X_1]^{-1}$$

$$R_{12} = -(X_1^T V^{-1} X_1)^{-1}(X_1^T V^{-1} X_2)[(X_2^T V^{-1} X_2) - (X_2^T V^{-1} X_1)(X_1^T V^{-1} X_1)^{-1}(X_1^T V^{-1} X_2)]^{-1}$$

$$R_{21} = -(X_2^T V^{-1} X_2)^{-1}(X_2^T V^{-1} X_1)[(X_1^T V^{-1} X_1) - (X_1^T V^{-1} X_2)(X_2^T V^{-1} X_2)^{-1}(X_2^T V^{-1} X_1)]^{-1}$$

$$R_{22} = [(X_2^T V^{-1} X_2) - (X_2^T V^{-1} X_1)(X_1^T V^{-1} X_1)^{-1}(X_1^T V^{-1} X_2)]^{-1}.$$

Therefore

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} .$$

Hence

$$(7.3) \qquad V(\hat{\beta}_1) = \frac{1}{[(X_1^T V^{-1} X_1) - (X_1^T V^{-1} X_2)(X_2^T V^{-1} X_2)^{-1}(X_2^T V^{-1} X_1)]}$$

$$\geq \frac{1}{(X_1^T V^{-1} X_1)} = \frac{1}{C_1^2} ,$$

Since $[(X_1^T V^{-1} X_1) - (X_1^T V^{-1} X_2)(X_2^T V^{-1} X_2)^{-1}(X_2^T V^{-1} X_1)] > 0$

and $(X_1^T V^{-1} X_2)(X_2^T V^{-1} X_2)^{-1}(X_2^T V^{-1} X_1) > 0.$

To show $V(\hat{\beta}_2) \quad \frac{1}{C_2^2}$ is shown similarly. The equality in (7.3) is satisfied whenever $X_1^T V^{-1} X_2 = 0.$

COROLLARY 7.1.1 For the special case of Theorem 7.1 where $V = I$, we have the condition $X_i^t X_i = C_i^2$, and the optimum choice of combinations (rows of X) is when $X_i^t X_j = 0$, that is, the columns of X are orthogonal. The proof follows from Theorem 7.1.

THEOREM 7.2 Given the linear model described above (Case I) where $V = \sigma^2 I$, show that the choice of the x_i, restricting the x_i to $-1 \leq x_i \leq 1$, which will give the minimum value for $\text{cov}(\hat{\beta})$ is $n/2$ choices of $x_i = -1$ and $n/2$ choices of $x_i = +1$.

Proof:

$$X^tX = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

$$(7.4) \qquad (X^tX)^{-1}\sigma^2 = \text{cov}(\hat{\beta}) = \begin{bmatrix} \dfrac{\sigma^2 \Sigma x_i^2}{n\Sigma(x_i - \bar{x})^2} & \dfrac{-\sigma^2 \Sigma x_i}{n\Sigma(x_i - \bar{x})^2} \\ \dfrac{-\sigma^2 x_i}{n\Sigma(x_i - \bar{x})^2} & \dfrac{+\sigma^2 n}{n\Sigma(x_i - \bar{x})^2} \end{bmatrix}$$

If we denote the columns of X by X_1 and X_2, and recall from Corollary 7.1.1 that X_1 and X_2 need to be orthogonal in order to obtain the minimum variance, we see that the terms off the diagonal in (7.4) must be zero. In order to make these terms zero, we must set $\Sigma x_i = 0$. In order to minimize $\text{Var}(\hat{\beta}_1)$, note that if $\Sigma x_i = 0$, then $\text{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{n}$, which does not depend on the values of the x_i. Finally, Since $\text{Var}(\hat{\beta}_2) = \dfrac{\sigma^2}{x_i^2}$ if $\Sigma x_i = 0$, we see that the Σx_i should be chosen so as to make x_i^2 as large as possible.

One method of forcing the x_i to sum to zero is to choose $x_{i+1} = -x_i$ for $i = 1,3,\ldots,n-1$. If the sum of the x_i^2 is to be as

large as possible, and if $-1 \leq x_i \leq +1$, the n we need to choose $n/2$ of the $x_i = 1$ and $n/2$ of the $x_i = -1$. This implies that $\text{var}(\beta_2) = \frac{\sigma^2}{n}$. We can reach the same conclusion in another manner. Consider the following example where $n = 4$.

$$\text{Choose } x_1 = 1/a_1 \text{ where } a \geq 1$$

$$x_2 = -x_1$$

$$x_3 = 1/a_2$$

$$x_4 = -x_3$$

Then

$$X = \begin{bmatrix} 1 & 1/a_1 \\ 1 & -1/a_1 \\ 1 & 1/a_2 \\ 1 & -1/a_2 \end{bmatrix}$$

$$X^tX = \begin{bmatrix} 4 & 0 \\ 0 & 2(1/a_1^2 + 1/a_2^2) \end{bmatrix}$$

$$\sigma^2(X^tX)^{-1} = \begin{bmatrix} \frac{\sigma^2}{4} & 0 \\ 0 & \frac{\sigma^2}{2}\left(\frac{(a_1^2)(a_2^2)}{(a_1^2)+(a_2^2)}\right) \end{bmatrix}$$

To minimize the $V(\beta_2)$, the term $\frac{a_1^2 a_2^2}{a_1^2 + a_2^2}$ needs to be made as small as possible. This can be accomplished only if $a_1 = a_2 = 1$. As a result, the matrix X becomes

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}$$

We now develop a technique for choosing the design matrix X whenever V is given. Let $V^{-1} = R = (r_{ij})$. Then write $Cov(\hat{\beta})$ as follows

$$Cov(\hat{\beta}) = \sigma^2(X^TRX)^{-1} = \sigma^2 \frac{\begin{bmatrix} \sum_i \sum_j ryX_iX_j & -\sum_i \sum_j X_i r_{ij} \\ -\sum_i \sum_j X_i r_{ij} & \sum \sum r_{ij} \end{bmatrix}}{(\sum \sum r_{ij})(\sum \sum r_{ij}X_iX_j) - (\sum \sum X_i r_{ij})^2}$$

By Theorem 2, we need to choose the off diagonal elements to be zero to insure $V(\hat{\beta}_1)$ and $V(\hat{\beta}_2)$ to be a minimum. Therefore $\sum \sum r_{ij}X_i = 0$ which implies $X_1^TRX_2 = 0$ since $X_1 = (1,1,\ldots,1)^T$ and $X_2 = (x_1, x_2, \ldots, x_n)^T$. The problem has now reduced to finding a vector X_2 such that $X_1^TRX_2 = 0$ and $V(\hat{\beta}_1)$ and $V(\hat{\beta}_2)$ are minimum. The above covariance matrix becomes

$$Cov(\hat{\beta}) = \begin{bmatrix} \frac{\sigma^2}{\Sigma\Sigma r_{ij}} & \phi \\ \phi & \frac{\sigma^2}{\sum_i \sum_j r_{ij}} \end{bmatrix}$$

Thus to minimize $V(\hat{\beta}_2)$ choose X_2 such that $X_2^TRX_2$ is as large as possible and $X_1^TRX_2 = 0$.

7.1 References

[1] Rao, C. R., Linear Statistic Inference and Its Application, John Wiley & Sons, Inc., 1966.

Chapter VIII

ON LINEAR ESTIMATION WITH LINEAR CONSTRAINTS

In this chapter we will develop the best linear estimates of x in the linear model

$$y = Hx + v \tag{8.1}$$

under a set of linear constraints.

We will consider the constraint

$$Ax = T$$

where A is an $m \times n$ known matrix such that

$$r(A) = m < n$$

and T is a known $m \times 1$ vector of constants.

Also we will consider estimating x in (4.1) if the added information that

$$r = Sx + u$$

is known where r is $t \times 1$ vector, S is a $t \times n$ known matrix and u is $t \times 1$ error vector such that

$$E(u) = \emptyset$$

and

$$e(uu^T) = \Lambda.$$

Λ is positive definite. No restriction on the rank of S is required.

8.1 Deterministic Constraints

Suppose that in addition to the linear model (8.1) it is known that

$$Ax = T \tag{8.2}$$

where A is $m \times n$ known matrix such that the rank of A

$$r(A) = m < n$$

and T is a known $m \times 1$ vector of constants. Our purpose is to develop a best estimator for x.

THEOREM 8.1 The minimum variance unbiased estimator for x in (8.1) given (8.2) is

$$\tilde{x} = \hat{x} + (h^T R^{-1} h)^{-1} A^T [A(h^T R^{-1} h)^{-1} A^T]^{-1} (T - Ax)$$

where

$$\hat{x} = (h^T R^{-1} h)^{-1} h^T R^{-1} y.$$

Proof: It is true that the "pay-off" function

$$Q = (y - Hx)^T R^{-1} (y - Hx)$$

yields the minimum variance estimate for x, that is

$$\frac{\partial Q}{\partial x} = -2H^T R^{-1}(y - Hx) = \emptyset$$

implies that

$$\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} y.$$

Consider the payoff function

$$Q' = (y - Hx)^T R^{-1}(y - Hx) + 2\lambda^T(T - Ax)$$

where λ is a $m \times 1$ vector of Lagrangian multipliers. Then

$$\frac{\partial Q'}{\partial x} = -2H^T R^{-1}(y - Hx) - 2A^T\lambda \equiv \emptyset$$

(8.3)

$$\frac{\partial Q'}{\partial \lambda} = T - Ax = \emptyset$$

gives a necessary condition for Q' to be a minimum. This implies

$$A^T\lambda = (H^T R^{-1}H)x - H^T R^{-1}y$$

which implies that the estimate

$$(H^T R^{-1}H)\,\tilde{x} = H^R R^{-1}y + A^T\lambda$$

or

$$\tilde{x} = \hat{x} + (H^T R^{-1}H)^{-1}A^T\lambda. \tag{8.4}$$

If

$$\lambda = [A(H^T R^{-1}h)^{-1}A^T]^{-1}(T - A\hat{x}), \tag{8.5}$$

then (8.4) and (8.5) form a minimum variance solution to the system of equations in (8.3). It is important to note that

$$\begin{aligned} T - A\tilde{x} &= T - A[\hat{x} + (H^T R^{-1}H)^{-1}A^T\lambda] \\ &= T - A\hat{x} - A(H^T R^{-1}H)A^T\lambda \\ &= T - A\hat{x} - (T - A\hat{x}) \\ &= \emptyset, \end{aligned}$$

as was required.

Also one notes that

$$E\tilde{x} = E\hat{x} + (H^T R^{-1} H)^{-1} E\lambda.$$

But

$$E\hat{x} = x$$

and

$$E\lambda = [A(H^T R^{-1} h)^{-1} A^T]^{-1} (T - AEx)$$

$$= [A(H^T R^{-1} h)^{-1} A^T]^{-1} (T - Ax)$$

$$= \emptyset$$

by (8.2).

The covariance matrix $C(\tilde{x})$ is obtained as follows:

$$C(\tilde{x}) = C\{\hat{x} + (h^T R^{-1} h)^{-1} A^T [A(h^T R^{-1} h)^{-1} A^T]^{-1} (T - A\hat{x})\}$$

$$= C\{(I - (h^T R^{-1} h)^{-1} A^T [A(h^T R^{-1} h)^{-1} A^T]^{-1} A)\hat{x}\}$$

$$= B(h^T R^{-1} h)^{-1} B^T$$

where

$$B = (I - (h^T R^{-1} h)^{-1} A^T [A(h^T R^{-1} h)^{-1} A^T]^{-1} A.$$

On reducing $C(x)$ to a simple form we find that

$$C(\tilde{x}) = (h^T R^{-1} h)^{-1} - (h^T R^{-1} h)^{-1} A^T [A(h^T R^{-1} h)^{-1} A^T]^{-1} A(h^T R^{-1} h)^{-1}$$

And again, added information leads to a reduction in the variance of the estimators.

8.2 Linear Constraints with Additive Random Components

Suppose the rank of h is n and that we know the additional information about the $n \times 1$ vector of unknown parameters, that is

$$r = Sx + u \tag{8.6}$$

where r is a $t \times 1$ vector, S is a $t \times n$ known matrix and u is a $t \times 1$ error vector such that $E(u) = \emptyset$ and $E(uu^T) = \Lambda$, a positive definite matrix.

Suppose the elements of u are independent of the elements of v in (1). We can then combine (8.1) and (8.6) to obtain a new linear model, that is

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} h \\ s \end{bmatrix} x + \begin{bmatrix} v \\ u \end{bmatrix} \tag{8.7}$$

where

$$E \begin{bmatrix} v \\ u \end{bmatrix} = \emptyset; \quad E \begin{bmatrix} v \\ u \end{bmatrix} (v,u) = \begin{bmatrix} R & \emptyset \\ \emptyset & \Lambda \end{bmatrix}.$$

The classical Gauss-Markov Theorem yields the estimator using (8.7) as the linear model as

$$\tilde{x} = [h^T R^{-1} h + s\Lambda^{-1} s]^{-1} [h^T R^{-1} y + s^T \Lambda^{-1} r]$$

or

$$\tilde{x} = \hat{x} + [h^T R^{-1} h + s^T \Lambda^{-1} s]^{-1} \, s^T \Lambda^{-1} [r - s\hat{x}].$$

the desired result.

Chapter IX

ON LINEAR ESTIMATION WITH INEQUALITY CONSTRAINTS

This chapter has its purpose to specify a general framework for combining prior and sample information in the linear model when the prior information consists of linear inequality or both equality and inequality restraints on the individual coefficients or combinations thereof. Under this specification, the estimation of the parameters of the linear model if formulated as a problem of minimizing a quadratic form subject to a set of linear equalities and inequalities (i.e., a typical quadratic programming problem) and an algorithm is specified which may be used to efficiently solve the nonlinear programming problem. The properties of the restricted estimator are discussed and the formulation is extended to cover a set of linear regression equations.

9.1 The Basic Model

The sample observations are assumed to be generated by the following linear model:

$$y = X\beta + u, \tag{9.1}$$

$$E(u) = 0, \tag{9.2}$$

$$E(uu') = \sigma^2 I \tag{9.3}$$

where y is a (Tx1) vector of observations, X is a (TxK) matrix on nonstochastic elements that are fixed in repeated samples and have rank K, and u is a vector of random disturbances which are assumed to have zero mean, (9.2), and constant (finite) variance, and be uncorrelated, (9.3). I is a unit matrix of order T, and β is a vector of unknown parameters to be estimated.

Given this specification for statistical model, when only the sample information is used, the ordinary least-squares estimator $\hat{\beta}$ is obtained by solving the following problem:

To find the vector $\hat{\beta}$ that minimizes

$$u'u = (y - X\beta)'(y - X\beta). \tag{9.4}$$

Setting the derivative of $u'u$ with respect to β equal to zero, yields, for the minimizing value, the least-squares estimator

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{9.5}$$

which can be shown to have a minimum variance within the class of all unbiased estimators which are a linear function of y.

9.2 General Restricted Least-Squares

We now reformulate the above problem to include the following inequality restraints:

$$\begin{bmatrix} I_1 \\ -I_1 \end{bmatrix} [\beta_1] \leq \begin{bmatrix} r_1^u \\ -r_1^s \end{bmatrix} \text{ for } r_1^u > 0 \text{ and } r_1^s \geq 0 \text{ or } 0 \leq r_1^s \leq \beta_1 \leq r_1^u, \tag{9.5a}$$

$$(9.5b) \qquad \begin{bmatrix} I_2 \\ -I_2 \end{bmatrix} [\beta_2] \leq \begin{bmatrix} r_2^u \\ -r_2^s \end{bmatrix} \text{ for } r_2^u \leq 0 \text{ and } r_2^s < 0 \text{ or } r_2^s \leq \beta_2 \leq r_2^u \leq 0,$$

$$(9.5c) \qquad \begin{bmatrix} I_3 \\ -I_3 \end{bmatrix} [\beta_3] \leq \begin{bmatrix} r_3^u \\ -r_3^s \end{bmatrix} \text{ for } r_3^u > 0 \text{ and } r_3^s < 0 \text{ or } r_3^s \leq \beta_3 \leq r_3^u \text{ and } r_3^s \leq 0 \leq r_3^u ,$$

$$(9.5d) \qquad \begin{bmatrix} S_1 & S_2 & S_3 \\ -S_1 & -S_2 & -S_3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \leq \begin{bmatrix} r_4^u \\ -r_4^s \end{bmatrix} \text{ for } r_4^u , r_4^s \gtrless 0 \text{ or } r_4^s \leq S\beta \leq r_4^u ,$$

$$(9.5e) \qquad [R_1 \quad R_2 \quad R_3] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = r_5 \text{ for } r_5 \gtrless 0,$$

where r_i^u and r_i^s for $i = 1,2,3$, are known vectors of upper and lower bound constraints for the unknown coefficients in the i^{th} set β_i, and I_i for $i = 1,2,3$, is the identity matrix with rank corresponding to the number of elements included in the parameter vector, β_i. Thus for the inequality constraint (9.5a) the unknown elements of the parameter vector β_i are restricted within a non-negative interval. Likewise for inequality constraints (9.5b) and (9.5c), the unknown elements β_2 are restricted to the non-positive interval and the unknown elements of β_3 may range over an interval of positive and negative values. Constraint

(9.5d) reflects the interval inequality specification for some linear combination of the unknown coefficients β_i with S representing a constant coefficient matrix reflecting linear combinations, ratios, etc. of the parameter vector β_i and r_4^s and r_4^u again represent a known vector of upper and lower bound constraints. Constraint (9.5e) reflects the equality specification on individual and/or a combination of unknown parameters β_i. In the formulation to follow this restraint will be handled as a special case of (9.5d) which relates to the situation when the upper and lower bound constraints are equal.

By extending the prior information on the parameter set β to include the inequality constraint possibilities (9.5a) to (9.5d) we have the following problem:

> To find the vector β^{**} that minimizes (9.4)
> subject to constraints (9.5a) through (9.5e)

9.3 General Restricted Formulation

To convert the problem of finding the vector β^{**} which minimizes (9.4) subject to the prior information contained in (9.5a) through (9.5e) to a quadratic programming problem, it is necessary to redefine the variables associated with the admissible non-positive coefficients so that the non-negativity requirements are fulfilled. In order to construct a feasible quadratic programming convention which does not conflict with statistical procedures:

Consistent with (9.4) and constraints (9.5a) through (9.5e) let the set of fixed X's be partitioned into X_1, X_2, X_3 where $[X_1, X_2, X_3] = [X]$ with the total number of variables equal to K and

$K = K_1+K_2+K_3$. With the subset X_1 associate the non-negative restricted vector of coefficients β_1 with the subset X_2 associate the non-positive vector of coefficients β_2, i.e., the $X_2'y$ vector is spanned by a negative extension of the vectors $(X_2'X_2)$ in the moment space. With the subset X_3 associate the vector of unrestricted sign parameters β_3. To handle the non-positive prior specification on β_2 and the negative possibility for β_3 and convert the expansion factors for the vectors of $(X_2'X_2)$ and $(X_3'X_3)$ to non-negative numbers, let us use the programming convention of treating X_2 as $-X_2$ and augmenting X_3 to $(X_3, -X_3)$ and β_3 to $\begin{bmatrix}\beta_3 \\ {}_3\beta\end{bmatrix}$, where ${}_3\beta$ corresponds to the negative counterpart of β_3. Under this convention define the augmented matrix $[X_1 -X_2 X_3 -X_3]$ as $\tilde{X}$ and the augmented vector $[\beta_1' \beta_2' \beta_3' {}_3\beta']'$ as $\tilde{\beta}$. Given this specification the problem may be formulated as:

To find the vector $\tilde{\beta}$ that maximizes the quadratic function

$$-\tilde{u}'\tilde{u} = -(y - \tilde{X}\tilde{\beta})'(y - X) = -y'y + 2\tilde{\beta}'\tilde{X}'y - \tilde{\beta}'\tilde{X}'\tilde{X}\tilde{\beta}$$

$$= -y'y + 2[\beta_1 \beta_2 \beta_3 {}_3\beta] \begin{bmatrix} X_1' \\ -X_2' \\ X_3' \\ -X_3' \end{bmatrix} [y]$$

(9.6)

$$- [\beta_1' \beta_2' \beta_3' {}_3\beta'] \begin{bmatrix} X_1'X_1 & X_1'(-X_2) & X_3'X_3 & X_1'(-X_3) \\ (-X_2)'X_1 & X_2'X_2 & (-X_2)'X_3 & X_2'X_3 \\ X_3'X_1 & X_3'(-X_2) & X_3'X_3 & X_3'(-X_3) \\ (-X_3)'X_1 & X_3'X_2 & (-X_3)'X_3 & X_3'X_3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ {}_3\beta \end{bmatrix}$$

$$(9.7)\quad \begin{bmatrix} I_1 & & & \\ -I_1 & & & \\ & I_2 & & \\ & -I_2 & & \\ & & I_3 & \\ & & & -I_3 \\ S_1 & -S_2 & S_3 & -S_3 \\ -S_1 & S_2 & -S_3 & S_3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ {}_3\beta \end{bmatrix} \leq \begin{bmatrix} r_1^u \\ -r_1^s \\ r_2^u \\ -r_2^s \\ r_3^u \\ -r_3^s \\ r_4^u \\ -r_4^s \end{bmatrix}$$

or

$$(9.7)\qquad \tilde{A\beta} \leq r,$$

and

$$(9.8)\qquad \tilde{\beta} \geq 0.$$

where $\tilde{X}^T\tilde{X}$ is $(K+K_3)x(K+K_3)$ positive semidefinite symmetric matrix and $\tilde{\beta}$ is a non-negative vector of unknown coefficients. Equality constraint (9.5e) is handled through the inequality constraint (9.5d).

In this form the specification (9.6), (9.7) and (9.8) reflects a typical quadratic programming problem, i.e., find the non-negative values of the $\tilde{\beta}$'s which minimize a quadratic function which satisfy given linear inequality constraints. By making use of the Kuhn-Tucker "equivalence theorem' for non-linear programming [11, 13, 19] and the

duality theorem of Dorn [5, 6, 19] for quadratic programs the following primaldual programming formulation results:

To maximize

$$(9.9) \qquad (\tilde{X}'y - \tilde{X}'\tilde{X}\beta)'\tilde{\beta} - \lambda' r = -(\beta' w + \lambda' v) \leq 0$$

subject to

$$(9.10) \qquad A\tilde{\beta} \leq r \quad \text{or} \quad A\tilde{\beta} + v = r,$$

$$(9.11) \qquad A'\lambda + (\tilde{X}'\tilde{X})\tilde{\beta} \geq \tilde{X}y \quad \text{or} \quad A'\lambda + (\tilde{X}'\tilde{X})\tilde{\beta} - w = \tilde{X}y,$$

and

$$(9.12) \qquad \tilde{\beta}, \lambda \geq 0 \quad \text{or} \quad \tilde{\beta}, X, v, w, \geq 0,$$

where λ is a vector of dual variables pertaining to the constraints and v and w are vectors of artificial variables for transforming (9.10) and (9.11) into equality systems. This problem, (9.9) through (9.12), is solvable by use of the standard simplex version of the quadratic programming algorithm developed by Wolfe [30]. The characteristics of this algorithm for our problem are reflected in the tableau presented in Table 1.

TABLE 1

SIMPLEX TABLEAU FOR GENERALIZED RESTRICTED LEAST-SQUARES ESTIMATORS

C			0	0	0	0
		P_o	λ	$\tilde{\beta} \geq 0$	$v \geq 0$	$w \geq 0$
-M -M	Z_1 Z_2	r $\tilde{X}'y$	A'	A $\tilde{X}'\tilde{X}$	I	$-I$

In Table 1 the I's are identity matrices and M is any positive real number attached to all artificial variables z_i of the initial basic solutions.

9.4 A Non-Linear Estimator

One can formulate a non-linear estimator for estimating a parameter vector X in the linear model under the constraints that

$$b_i \leq X_i \leq a_i \qquad i = 1,2,\ldots,n$$

by applying the Bayesian technique.

Suppose X has the multivariate uniform distribution and suppose that Y given X has the multivariate normal distribution. Hence we may write the respective density functions

$$h(Y/X) = \frac{1}{(2\pi)^{n/2} |R|^{1/2}} \exp\left(-\frac{1}{2}(Y-HX)^T R^{-1}(Y-HX)\right) \tag{9.13}$$

and

$$g(X) = \frac{1}{(a_1+b_1)(a_2+b_2)\ldots(a_p+b_p)} \tag{9.14}$$

where if $X = (x_i)$ then $-b_i \leq x_i \leq a_i$.

The Bayesian estimator for X, say $\hat{X}$ is given by

$$\hat{X} = E(X/Y). \tag{9.15}$$

The joint probability density of (Y,X) is

$$f(Y,X) = h(Y/X)\, g(X) \tag{9.16}$$

and the conditional density of X given Y is

$$g(X/Y) = f(Y,X)/h(x) \tag{9.17}$$

where

$$h(Y) = \int_{-b_p}^{a_p} \cdots \int_{-b_1}^{a_1} f(Y,X)dx_1\, dx_2 \cdots dx_p \tag{9.18}$$

$$= \frac{1}{(a_1+b_1)(a_2+b_2)\cdots(a_p+b_p)(2\pi)^{n/2}\,|R|^{1/2}}$$

$$\int_{-b_p}^{a_p} \cdots \int_{-b_1}^{a_1} \exp\left(-\tfrac{1}{2}(Y-HX)^T R^{-1}(Y-HX)\right)dx_1\, dx_2 \cdots dx_p.$$

Then it follows that

$$g(X,Y) = \frac{1}{(a_1+b_1)(a_2+b_2)\cdots(a_p+b_p)(2\pi)^{n/2}\,|R|^{1/2}} \tag{9.19}$$

$$\exp\left(-\tfrac{1}{2}(Y-HX)^T R^{-1}(Y-HX)\right) \div$$

$$\frac{1}{(a_1+b_1)(a_2+b_2)\cdots(a_p+b_p)(2\pi)^{n/2}\,|R|^{1/2}}$$

$$\int_{-b_p}^{a_p} \cdots \int_{-b_1}^{a_1} \exp\left(-\tfrac{1}{2}(Y-HX)^T R^{-1}(Y-HX)\right)dx_1\, dx_2 \cdots dx_p.$$

From (9.18) and (9.19) it follows that

$$X = \frac{\int_{-b_p}^{a_p} \cdots \int_{-b_1}^{a_1} X \exp(-\frac{1}{2}(Y-HX)^T R^{-1}(Y-HX))dx_1\, dx_2 \cdots dx_p}{\int_{-b_p}^{a_p} \cdots \int_{-b_1}^{a_1} \exp(-\frac{1}{2}(Y-HX)^T R^{-1}(Y-HX))dx_1\, dx_2 \cdots dx_p}. \tag{9.20}$$

Since $-b_i \leq x_i \leq a_i$ equation (9.20) implies that $-b_i \leq \hat{x}_i \leq a_i$.

Unfortunately we required normality assumptions on the observations to obtain $\hat{X} = (\hat{x}_i)$ however we may study $\hat{X}$ as defined by (9.20) as a non-linear estimator of X when X is not a random variable and Y is not normally distributed.

The properties of such an estimator have not been obtained at this time.

9.5 References

[1] Aitken, A. C., "On Least Squares and Linear Combination of Observations," Proceedings of the Royal Society of Edinburgh, Vol. 55, 1934-1935, pp. 42-8.

[2] Ashar, V., and Wallace, T. D., "A Sampling Study of Minimum Absolute Deviations Estimator," Operations Research, Vol. 11, Sept.-Oct. 1963, pp. 747-58.

[3] Chipman, J. S. and Rao, M. M, "The Treatment of Linear Restrictions in Regression Analysis," Econometrica, Vol. 32, No. 1-2, Jan.-April 1964, pp. 198-209.

[4] Cramer, H.,"Mathematical Methods of Statistics," Princeton, Princeton University Press, 1946, pp. 247-8.

[5] Dantzig, G. B. and Orden, A., "Duality Theorems," RAND Report R.M.-1265, The RAND Corp., Santa Monica, California, Oct. 1953.

[6] Dorn, W. S., "Duality in Quadratic Programming," Quarterly of Applied Mathematics, Vol. 18, 1960, pp. 155-62.

[7] Durbin, J., "A Note on Regression When There is Extraneous Information About One of the Coefficients," Journal of the American Statistical Assn., Vol. 48, Dec. 1953, pp. 799-808.

[8] Fisher, W. D., "A Note on Curve Fitting with Minimum Deviations by Linear Programming," Journal of the American Statistical Assn., Vol. 56, 1961, pp. 359-62.

[9] Goodman, L. A., "A Further Note on Miller's Finite Markov Processes in Psychology," Psychometrika, Vol. 18, 1953, pp. 245-8.

[10] Haavelmo, T., "The Statistical Implications of a System of Simultaneous Equations," Econometrica, Vol.11, 1943, pp. 1-12.

[11] Karlin, S., "Mathematical Methods and Theory in Games, Programming and Economics," London, Addison-Wesley, 1959, pp. 199-242.

[12] Karst, O. J., "Linear Curve Fitting Using Least Deviations," Journal of the American Statistical Assn., Vol. 53, 1958, pp. 118-32.

[13] Kuhn, H. and Tucker, A., "Non-Linear Programming," Proceedings of the Second Berkeley Symposium, J. Neyman (ed.), Berkeley, The University of California Press, 1951, pp. 481-92.

[14] Lee, T. C., Judge, G. G., and Takayama, T., "On Estimating the Transitional Probabilities of a Markov Process," Journal of Fram Economics, Vol. 47, 1965, pp. 742-62.

[15] Madansky, A, "Least Squares Estimation in Finite Markov Processes," Psychometrika, Vol. 24, 1959, pp. 137-44.

[16] Meyer, J. and Glauber, R. G., "Investment Decisions, Economic Forecasting and Public Policy," Boston, Harvard University, 1964.

[17] Miller, G. A., "Finite Markov Processes in Psychology," Psychometrika, Vol. 17, 1952, pp. 149-67.

[18] Raiffia, A. and Schlaifer, R., "Applied Statistical Decision Theory," Boston, Harvard University, 1961.

[19] Saaty, T. L. and Bram, J., "Non-Linear Mathematics," New York, McGraw-Hill Book Co., 1964, pp. 113-33.

[20] Takayama, T., and Judge, G. G., "Equalibrium Among Spatially Separated Markets: A Reformulation," Econometrica, Vol. 32, 1964, pp. 510-24.

[21] Theil, H., and Goldberger, A. S., "On Pure and Mixed Statistical Estimation in Economics," International Economic Review, Vol. 2, January 1961, pp. 65-78.

[22] Theil, H., "On the Use of Incomplete Prior Information in Regression Analysis," Journal of the American Statistical Assn., Vol. 58, June 1963, pp. 401-14.

[23] Theil, H., Economic Forecasts and Policy, Second Edition, Amsterdam, North Holland, 1961, pp. 331-3.

[24] Tintner, G., "Stochastic Linear Programming With Applications to Agricultural Economics," Proceedings of the Second Symposium in Linear Programming, Washington, D. C., pp. 197-228.

[25] Telser, L. G., "Least Squares Estimates of Transition Probabilities,' Chapter in the book Measurement in Economics, Stanford, Stanford University Press, 1963, pp. 272-92.

[26] Tiano, G. C. and Zellner, A., "Bayes's Theorem and the Use of Prior Knowledge in Regression Analysis," Biometrika, Vol. 51, No. 1 and 2, 1964, pp. 219-30.

[27] Tinter, G., Econometrics, New York, John Wiley and Sons, Inc., 1952, pp. 89-91

[28] Wagner, H. M., "Linear Programming for Regression Analysis," Journal of the American Statistical Assn., Vol. 54, 1959, pp. 206-12.

[29] Wilks, S. S., Mathematical Statistics, Princeton University Press, 1945, p. 174.

[30] Wolfe, P., "The Simplex Method for Quadratic Programming," Econometrica, Vol. 27, 1959, pp. 382-98.

[31] Zellner, A., "Linear Regression With Inequality Constraints on the coefficients," Mimeographed Report 6109 of The International Center for Management Science, 1961.

[32] Zellner, A., "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Assn., Vol 57, 1962, pp. 348-68.

[33] Zellner, A. and Theil, H., "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," Econometrica, Vol. 30, 1962, pp. 54-78.

Chapter X

ON RECURSIVE ESTIMATION WHEN THE COVARIANCE MATRIX IS UNKNOWN

10.1 The Estimators

It is well-known that the minimum variance linear unbiased estimate for the unknown parameter vector β in the linear model

$$Y_N = X_N\beta + e_N \tag{10.1}$$

is

$$\hat{\beta}_N = (X_N^T V_N^{-1} X_N)^{-1} X_N^T V_N^{-1} Y_N \tag{10.2}$$

where

Y_N is Np X 1 vector of observations,

β_N is n X 1 vector of parameters to be estimated.

X_N is Np X n matrix of known real numbers, and

e_N is Np X 1 vector of random errors

such that

$$Ee_N = \phi \tag{10.3}$$

$$Ee_N e_N^T = V_N \tag{10.4}$$

where

V_N is a Np X Np positive definite covariance matrix.

Clearly, in order to use the estimator $\hat{\beta}$ one must know V_N. Unfortunately, this is not usually the practical situation. One alternative is immediate. Instead of (10.2) use the estimator

$$\beta_N^* = (X_N^T X_N)^{-1} X_N^T Y_N \tag{10.5}$$

which minimizes the sum of squared errors,

$$e_N^T e_N \tag{10.6}$$

We will limit the investigation here to the case where V_N is block diagonal and each block is a $p \times p$ submatrix V, that is

$$V_N = \begin{bmatrix} V & \emptyset & . & . & . & \emptyset \\ \emptyset & V & . & . & . & \emptyset \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ \emptyset & \emptyset & . & . & . & V \end{bmatrix} \tag{10.7}$$

This case is important in orbit determination problems in which

$$Y_N = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix} \quad ; \qquad X_N = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_N \end{bmatrix}$$

where the sequence $\{y_i;\ i = 1, 2, \ldots, N\}$ is a sequence of $p \times 1$ observations. The vector Y_1 is the data obtained at time t_1, Y_2 the data vector at time t_2, etc. where $t_1 < t_2 < \ldots < t_N$. The estimation must be done recursively. The recursive forms of (10.2) and (10.5) are

$$\hat{\beta}_N = \hat{\beta}_{N-1} + [X_{N-1}^T V_{N-1}^{-1} X_{N-1} + x_N^T V^{-1} x_N]^{-1} x_N^T V^{-1} [y_N - x_N \hat{\beta}_{N-1}] \tag{10.8}$$

and

$$\beta_N^* = \beta_{N-1}^* + [X_{N-1}^T X_{N-1} + x_N^T x_N]^{-1} x_N^T [y_N - x_N \beta_{N-1}^*] \tag{10.9}$$

where

$$\begin{aligned} [X_N^T V^{-1} X_N]^{-1} &= [X_{N-1}^T V_{N-1}^{-1} X_{N-1} + x_N^T V^{-1} x_N]^{-1} \\ &= [X_{N-1}^T V_{N-1}^{-1} X_{N-1}]^{-1} - [X_{N-1}^T V_{N-1}^{-1} X_{N-1}]^{-1} x_N^T \\ &\quad [V + x_N (X_{N-1}^T V_{N-1}^{-1} X_{N-1})^{-1} x_N^T]^{-1} x_N [X_{N-1}^T V_N^{-1} X_{N-1}] \end{aligned} \tag{10.10}$$

The recursive form of $[X_N^T X_N]^{-1}$ can be obtained by letting $V = I$ in (10.10).

Other estimators that are recursive in nature which may be "good' estimators are

$$\tilde{\beta}_N = [\sum_{\alpha=1}^{N} x_\alpha^T S_\alpha^{-1} x_\alpha]^{-1} [\sum_{\alpha=1}^{N} x_\alpha^T S_\alpha^{-1} y_\alpha] \tag{10.11}$$

where

$$S_\alpha = \sum_{j=1}^{\alpha} (y_j - x_j \beta_j^*)(y_j - x_j \beta_j^*)^T / \alpha \tag{10.12}$$

and

$$\beta_N = \left(\sum_{\alpha=1}^{N} \chi_\alpha T_{\alpha-1}^{-1}\right)^{-1}\left[\sum_{\alpha=1}^{N} \chi_\alpha T_{\alpha-1}^{-1} y_\alpha\right] \tag{10.13}$$

where

$$T_\alpha = \sum_{j=1}^{\alpha} (y_j - x_j \tilde{\beta}_{j-1})(y_j - x_j \tilde{\beta}_{j-1})^T/\alpha \tag{10.14}$$

given that β_0 is the best approximation for β at time t_0.

If a recursive scheme is not necessary then one might conjecture that the estimator defined by (10.11) with the modification that S_α be replaced by

$$S_\alpha = \sum_{j=1}^{N} (y_j - x_j \beta_N^*)(y_j - x_j \beta_N^*)^T/N \tag{10.15}$$

for all α or

$$S_\alpha = \sum_{j=1}^{N} (y_j - x_j \hat{\beta}_{j-1})(y_j - x_j \hat{\beta}_{j-1})^T/N \tag{10.16}$$

for all .

The purpose of this chapter is to study the properties of the estimators defined by (10.11) through (10.14).

10.2 Properties of the Estimators

The following well-known [1] lemma will be used in establishing the properties of the estimators defined by (10.11) and (10.13) and the statistics (10.12) and (10.14).

LEMMA 10.1 Let X and Y be a n x 1 and m x 1 vector of random variables, respectively. Then

(10.15) $$E(Y) = E\{E(Y|X)\}$$

Proof: Let f(X,Y) denote the joint probability density function whose first moments exist. By definition of expectation

$$E(Y) = \int_x \int_y y f_{XY}(x,y)\, dx\, dy$$

$$= \int_x \int_y y f_{Y/X}(y)\, f_X(x)\, dx\, dy$$

$$= \int_x E(Y|X)\, f_X(x)\, dx$$

$$= E\{E(Y|X)\},$$

the desired result.

Consider then $E(\tilde{\beta}_N)$, that is

$$E(\tilde{\beta}_N) = E\,[\sum_{\alpha=1}^{N} x_\alpha s_\alpha^{-1} x_\alpha]^{-1}\,[\sum_{\alpha=1}^{N} x_\alpha^T s_\alpha^{-1} y_\alpha]$$

$$= E\{[\sum_{\alpha=1}^{N} x_\alpha s_\alpha^{-1} x_\alpha]^{-1}\,[\sum_{\alpha=1}^{N} x_\alpha^T s_\alpha^{-1} (Ey_\alpha|s_\alpha]\}$$

$$= E\{[\sum_{\alpha=1}^{N} x_\alpha s_\alpha^{-1} x_\alpha]^{-1}\,[\sum_{\alpha=1}^{N} x_\alpha s_\alpha^{-1} x_\alpha \beta]$$

$$= E\beta$$

$$= \beta.$$

Similarly, $E(\bar{\beta}_N) = \beta$. Hence both $\hat{\beta}_N$ and $\bar{\beta}_N$ are unbiased estimators for the parameter vector β.

The expected value of the statistic defined by (10.12) is

$$Es_\alpha = E\{\sum_{j=1}^{\alpha}(y_j - x_j\beta_j^*)(y_j - x_j\beta_j^*)^T|\alpha\}$$

$$= \frac{1}{\alpha} E\{\sum_{j=1}^{\alpha}[(y_j - x_j\beta)-x_j(\beta_j^* - \beta)][(y_j - x_j\beta)-x_j(\beta_j^* - \beta)]^T\}$$

$$= V - \frac{2}{\alpha}\sum_{j=1}^{\alpha}E(\beta_j^*-\beta)(y_j-x_j\beta)^T + \frac{1}{\alpha}\sum_{j=1}^{\alpha}x_jE(\beta_j^*-\beta)(\beta_j^*-\beta)^Tx_j^T.$$

Now

$$E(\beta_j^* - \beta)(y_j - x_j\beta)^T$$

$$= E[(\beta_{j-1}^*-\beta)(y_j-x_j\beta)^T]+E[(X_{j-1}^TX_{j-1}+x_j^Tx_j)^{-1}$$

$$x_j^T(y_j-x_j\beta_{j-1}^*)(y_j-x_j\beta_{j-1}^*)^T].$$

Let $K_j = (X_{j-1}^TX_{j-1} + x_j^Tx_j)^{-1}x_j^T$, then

$$E(\beta_j^*-\beta)(y_j-x_j\beta)^T = 0 + K_jE(y_j-x_j\beta)(y_j-x_j\beta)^T - K_jx_jE(\beta_{j-1}^*-\beta)(y_j-x_j\beta)^T$$

$$= K_jV.$$

Also

$$E(\beta_j^*-\beta)(\beta_j^*-\beta)^T = (X_j^TX_j)^{-1}X_j^TV_jX_j(S_j^TS_j)^{-1}.$$

Therefore

$$ES_\alpha = V - \frac{2}{\alpha}\sum_{j=1}^{\alpha}(X_{j-1}^T X_{j-1}+x_j^T x_j)^{-1}\, x_j^T V + \frac{1}{\alpha}\sum_{j=1}^{\alpha} x_j (X_j^T X_j)^{-1} X_j^T V_j X_j (X_j^T X_j)^{-1} x_j^T.$$

This shows the S_α is a biased estimate.

It remains to be determined the covariance matrix of the estimates (10.11) and (10.13). Since the expected value of the estimate (10.14) depends on the covariance of the estimate (10.11), this also remains to be done.

10.3 References

[1] Graybill, F. A., An Introduction to Statistical Models, Vol. 1, McGraw-Hill, Inc., 1961, p. 199.

Chapter XI

THE MAXIMUM LIKELIHOOD ESTIMATES OF THE MEAN VECTOR AND THE COVARIANCE MATRIX

11.1 Summary

The maximum likelihood estimators are derived for the mean vector μ and the covariance matrix R where the $p \times 1$ random vector X is distributed as

$$N(X; \mu, R) = \frac{1}{(2\pi)^{p/2}|R|^{1/2}} \exp -\frac{1}{2}(X - \mu)^T R^{-1}(X - \mu). \tag{11.1}$$

11.2 Preliminary Notions and Notation

Let $\{X_\alpha, \alpha = 1,2,\ldots,N\}$ represent a sample of N observations on X according to (11.1), where $N > p$.

$$L = \frac{1}{(2\pi)^{pN/2}|R|^{N/2}} \exp -\frac{1}{2} \sum_{\alpha=1}^{N} (X_\alpha - \mu)^T R^{-1}(X_\alpha - \mu). \tag{11.2}$$

Since the exponent is written in terms of R^{-1}, we shall find the maximum likelihood estimates of μ and R^{-1}. The following lemma will give the maximum likelihood estimate of R from the maximum likelihood estimate of R^{-1}, say ψ.

LEMMA 11.1 [1] Let $f(\theta)$ be a real-valued function defined on a certain set S and let ϕ be a single-valued function, with a single-valued inverse on S to some set S^*, that is, to each $\theta \in S$ there corresponds a unique $\theta^* \in S^*$ and conversely to each $\theta^* \in S^*$ there corresponds a unique $\theta \in S$.

Let

$$g(\theta^*) = f[\phi^{-1}(\theta^*)].$$

Then if $f(\theta)$ attains a maximum at $\theta = \theta_0$, $g(\theta^*)$ attains a maximum at $\theta^* = \theta_0^* = \phi(\theta_0)$. If the maximum of $f(\theta)$ at θ_0 is unique, so is the maximum of $g(\theta^*)$ at θ_0^*.

Other useful lemmas are

LEMMA 11.2 If $A = A^T$ then

$$\partial|A|/\partial a_{ii} = A_{ii}$$

$$\partial|A|/\partial a_{ij} = 2A_{ij}$$

where $A = \{a_{ij}\}$ and the minor of a_{ij} is A_{ij} (a scalar).

and

LEMMA 11.3 Let $X_1, X_2, \ldots, X_N$ be N (p-component) vectors and let $X = \sum_{\alpha=1}^{N} X_\alpha/N$. Then for any vector b,

$$\sum_{\alpha=1}^{N} (X_\alpha - b)(X_\alpha - b)^T \tag{11.4}$$

$$= \sum_{\alpha=1}^{N} (X_\alpha - X)(X_\alpha - X) + n(X - b)(X - b).$$

When we let $b = \mu$ in (11.4) and let

$$A = \sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})^T,$$

the quantity (11.4) can be written

$$\sum_{\alpha=1}^{N} (X_\alpha - \mu)(X_\alpha - \mu)^T = A + N(\bar{X} - \mu)(\bar{X} - \mu). \tag{11.6}$$

Using the result and the properties of the trace of a matrix ($\operatorname{tr} CD = \sum_{ij} c_{ij}d_{ji} = \operatorname{tr} DC$) we have

$$\begin{aligned} \sum_{\alpha=1}^{N} (X_\alpha - \mu)^T \psi (X_\alpha - \mu) &= \operatorname{tr} \sum_{\alpha=1}^{N} (X_\alpha - \mu)^T \psi (X_\alpha - \mu) \\ &= \operatorname{tr} \sum_{\alpha=1}^{N} \psi (X_\alpha - \mu)(X_\alpha - \mu)^T \\ &= \operatorname{tr} \psi A + \operatorname{tr} \psi N(\bar{X} - \mu)(\bar{X} - \mu)^T \\ &= \operatorname{tr} \psi A + N(\bar{X} - \mu)^T \psi (\bar{X} - \mu). \end{aligned} \tag{11.6}$$

Thus we can write $\log L$ where L is given by (11.1).

$$\begin{aligned} \log L = &- \frac{1}{2} pN \log (2\pi) + \frac{1}{2} N \log |\psi| \\ &- \frac{1}{2} \operatorname{tr} \psi A - \frac{1}{2} N(\bar{X} - \mu)^T \psi (\bar{X} - \mu). \end{aligned} \tag{11.7}$$

Since ψ is positive semidefinite, $N(\bar{X} - \mu)^T \psi (\bar{X} - \mu) \geq 0$ and 0 if $\mu = \bar{X}$. To maximize the second and third terms of (11.7) we use the following lemma

LEMMA 11.4 Let

$$f(c) = \frac{1}{2} N \log |c| - \frac{1}{2} \sum_{ij} c_{ij} d_{ji}$$

where $C = (c_{ij})$ is positive semidefinite and where $D = (d_{ij})$ is positive definite. Then the maximum of $f(c)$ is taken on at $C = ND^{-1}$ and the maximum is

$$f(ND^{-1}) = \frac{1}{2} pN \log N - \frac{1}{2} N \log |D| - \frac{1}{2} N.$$

Proof: We note that $f(c)$ tends to $-\infty$ if C approaches a singular matrix or if one or more elements of C approach ∞ and/or $-\infty$. Thus maxima of $f(c)$ are defined by setting equal to zero the derivatives with respect to the element of c. Using lemma 11.2 above, we find

$$\frac{\partial f}{\partial c_{kk}} = \frac{1}{2} \frac{N}{|c|} \frac{\partial |c|}{\partial c_{kk}} - \frac{1}{2} d_{kk} = \frac{1}{2} N \frac{\text{cof } c_{kk}}{|c|} - \frac{1}{2} d_{kk} \tag{11.8}$$

where $\text{cof } c_{kk}$ denotes the cofactors of c_{kk} in C. For $k \neq \ell$

$$\frac{\partial f}{\partial c_{k}} = N \frac{\text{cof } c_{kk}}{|c|} - d_{k\ell}, \tag{11.9}$$

since $c_{k\ell} = c_{\ell k}$. Setting $2\partial f/\partial c_{kk}$ and $\partial f/\partial c_{k\ell}$ equal to 0 and using the fact that $\text{cof } c_{k\ell}/|c|$ is the ℓ,k^{th} element of C^{-1}, we obtain $NC^{-1} = D$. Thus $C = ND^{-1}$. The value of maximum is

$$\begin{aligned} f(ND^{-1}) &= \frac{1}{2} N \log |ND^{-1}| - \frac{1}{2} \text{tr } ND^{-1}D \\ &= \frac{1}{2} N \log N^{p} |p^{-1}| - \frac{1}{2} \text{tr } NI \\ &= \frac{1}{2} Np \log N - \frac{1}{2} N \log |D| - \frac{1}{2} Np \end{aligned}$$

and the lemma is proved.

11.3 The Estimators

On applying lemma (11.3) to (11.4) we find that the maximum occurs at

$$\psi = \frac{1}{N} A^{-1} .$$

We assume that A is nonsingular (the probability is 0 of drawing a sample $N > p$ such that A is singular). Thus A^{-1} exists, and ψ is positive definite. Therefore $\mu = \bar{X}$ is the only value of μ to make the last term of (11.7) zero. Thus the maximum likelihood estimators for μ and ψ are $\hat{\mu} = \bar{X}$ and $\psi = NA^{-1}$.

To find the maximum likelihood estimator of R, we apply lemma 11.1, giving

$$\hat{R} = A/N,$$

which with

$$\hat{\mu} = \bar{X}$$

are the maximum likelihood estimators.

11.4 The Case When $\mu = H_\alpha X$ and Covariance Matrix is Known

The likelihood function is

$$L = \frac{1}{(2\pi)^{Np/2}} \frac{1}{|R|^{N/2}} e^{-\frac{1}{2} \sum_{\alpha=1}^{N} (Y_\alpha - H_\alpha X)^T R^{-1} (Y_\alpha - H_\alpha X)}.$$

Consider the natural log of L and take the partial derivative of

$\ell n\ L$ with respect to the vector X, that is

$$\frac{\partial \ell nL}{\partial X} = \sum_{\alpha=1}^{N} H_\alpha^T R^{-1}(Y_\alpha - H_\alpha X)2$$

Letting $\frac{\partial \ell nL}{\partial X} = 0$ and solving for $\tilde{X}$, the maximum likelihood estimator for X yields the minimum variance linear unbiased estimator for X, that is

$$\tilde{X} = (\sum_{\alpha=1}^{N} H_\alpha^T R^{-1} H_\alpha)^{-1} \sum_{=1}^{N} H_\alpha R^{-1} Y_\alpha .$$

A normality assumption is simply an unnecessary added restriction.

11.5 References

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., 1958, pp. 44-49.

Chapter XII

ON COMBINING UNBIASED VECTOR ESTIMATORS OF A VECTOR PARAMETER

12.1 Preliminary Concepts

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of $p \times 1$ vector parameter θ. The $p \times p$ covariance matrices of $\hat{\theta}_1$ and $\hat{\theta}_2$ will be denoted by R_1 and R_2; and the unbiased estimators of R_1 and R_2 be $\hat{R}_1$ and $\hat{R}_2$, respectively. We seek a linear combination of $\hat{\theta}_1$ and $\hat{\theta}_2$ which will be an unbiased estimator for θ and have a minimal covariance matrix in the following sense:

DEFINITION 1.1 The covariance matrix R is minimal if for any other covariance matrix Q, the matrix $Q - R$ is not negative semidefinite or negative definite.

DEFINITION 1.2 The covariance matrix R is strictly minimal if for any other covariance matrix Q, $Q - R$ is positive semidefinite or positive definite.

Let the combined estimator for θ be defined as

$$\hat{\theta} = A\hat{\theta}_1 + B\hat{\theta}_2 \tag{12.1}$$

where A and B are $p \times p$ matrices of real elements.

Since $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased, then $E\hat{\theta} = (A + B)\theta$ which implies that in order for $\hat{\theta}$ to be unbiased

$$(A + B) = I \tag{12.2}$$

where I is the $p \times p$ identity matrix.

The covariance matrix of $\hat{\theta}$ is

$$E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = AR_1A^T + AR_{12}B^T + BR_{21}A^T + BR_2B^T \tag{12.3}$$

where

$$R_{12} = E[(\hat{\theta}_1 - \theta)(\hat{\theta}_2 - \theta)^T]$$

$$R_{21} = E[(\hat{\theta}_2 - \theta)(\hat{\theta}_1 - \theta)^T]$$

$$R_1 = E[(\hat{\theta}_1 - \theta)(\hat{\theta}_1 - \theta)^T]$$

and $$R_2 = E[(\hat{\theta}_2 - \theta)(\theta_2 - \theta)^T]$$

Using the techniques from the calculus of variations and solving for the matrix A after equating the first variation of R with respect to A to zero in (12.3) under the constraints (12.2) one obtains

$$A = (R_2 - R_{21})[R_1 + R_2 - R_{21} - R_{12}] \tag{12.4a}$$

$$B = (R_1 - R_{12})[R_1 + R_2 - R_{21} - R_{12}] \tag{12.4b}$$

and finally

$$R = R_2 - (R_2 - R_{21})[R_1 + R_2 - R_{21} - R_{12}]^{-1}(R_2 - R_{21}) \tag{12.5a}$$

$$R = R_1 - (R_1 - R_{12})[R_1 + R_2 - R_{21} - R_{12}]^{-1}(R_2 - R_{21}) \tag{12.5b}$$

where we assume the existence of the inverse of

$$[R_1 + R_2 - R_{21} - R_{12}]^{-1}.$$

Usually, the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are uncorrelated; that is, $R_{12} = R_{21} = \Phi$, then (12.4) reduces to

$$A = R_2[R_1 + R_2]^{-1} = [R_1^{-1} + R_2^{-1}]R_1^{-1} = RR_1^{-1} \tag{12.6a}$$

$$B = R_1[R_1 + R_2]^{-1} = [R_1^{-1} + R_2^{-1}]R_2^{-1} = RR_2^{-1} \tag{12.6b}$$

where

$$R^{-1} = R_1^{-1} + R_2^{-1}. \tag{12.7}$$

These results are well-known [1] and are included here for sake of completeness. The covariance matrix R is strictly minimal when compared with the covariance matrix of any other linear combination estimator which is unbiased.

12.2 The Combined Estimator When the Covariance Matrices R_1 and R_2 Are Unknown

Let (MVN denotes multivariate normal)

$$\begin{aligned} X^{(1)} &\sim \text{MVN}\ (\theta, R_1) \\ X^{(2)} &\sim \text{MVN}\ (\theta, R_2) \end{aligned} \tag{12.8}$$

then

$$\bar{X}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} X_i$$

and

$$\bar{X}_2 = \frac{1}{N_2} \sum_{i=1}^{N_2} X_i$$

are unbiased estimators with covariance matrices $R_1/(N_1 - 1)$ and $R_2/(N_2 - 1)$, respectively. Let the random samples $\{X_i^{(1)}; i = 1,\ldots,N_1\}$ and $\{X_i^{(2)}; i = 1,\ldots,N_2\}$ be independent, then $\bar{X}_1$ and $\bar{X}_2$ are independent. It is well-known that $\bar{X}_j$ and S_j, where

$$S_j = \frac{1}{N_j - 1} \Sigma (X_i^{(j)} - \bar{X}_j)(X_i^{(j)} - \bar{X}_j)^T \tag{12.9}$$

are independent [2]. Hence $\bar{X}_1$, $\bar{X}_2$, S_1 and S_2 are mutually independent. Also we note that

$$ES_j = R_j \qquad j = 1,2.$$

Forming the estimator by substituting the maximum likelihood estimators S_j for R_j in (12.6) and (12.1) and taking the expectation it is found that the estimator

$$\tilde{\theta} = S_2[S_1 + S_2]^{-1}X_1 + S_1[S_1 + S_2]^{-1}X_2 \tag{12.10}$$

is unbiased. This follows directly

$$E\tilde{\theta} = E\{S_2(S_1 + S_2)^{-1}\}E\{X_1\} + E\{S_1[S_1 + S_2]^{-1}\}EX_2 = \theta.$$

We note since $(A + B) = I$, that $E\tilde{\theta} = \theta$ even if S_1 and S_2 are correlated and X_1 and X_2 remain uncorrelated with S_1 and S_2. However, difficulties will arise when correlation exists between the estimators and the estimates of their covariances.

For clarity consider the univariate case in which we wish to estimate σ from two separate sources. Let

$$x^{(1)} \sim N(\mu_1, \sigma^2)$$

$$x^{(2)} \sim N(\mu_2, \sigma^2)$$

and μ_1 and μ_2 are not known. Two unbiased estimators for σ^2 are

$$s_1^2 = \frac{1}{N_1 - 1} \sum_{\alpha=1}^{N_1} (X_\alpha^{(1)} - \bar{X}_1)^2$$

$$s_2^2 = \frac{1}{N_2 - 1} \sum_{\alpha=1}^{N_2} (X_\alpha^{(2)} - \bar{X}_2)^2.$$

It is well-known that

$$\frac{(N_1 - 1)s_i^2}{\sigma^2} \sim \chi^2(N_i - 1)$$

and it follows that

$$\text{Var } s^2 = \frac{2\sigma^4}{N_i - 1}$$

Using (12.1) and (12.6) it follows that the strictly minimal unbiased estimator for σ is

$$\theta = \frac{\frac{2\sigma^4}{N_2 - 1}}{\frac{2\sigma^4}{N_2 - 1} + \frac{2\sigma^4}{N_1 - 1}} s_1^2 + \frac{\frac{2\sigma^4}{N_1 - 1}}{\frac{2\sigma^4}{N_2 - 1} + \frac{2\sigma^4}{N_1 - 1}} s_2^2$$

or

$$\theta = \frac{N_1 - 1}{N_1 + N_2 - 2} s_1^2 + \frac{N_2 - 1}{N_1 + N_2 - 2} s_2^2 \tag{12.11}$$

a well-known result. It is important to note that it is not necessary to estimate the variances of s_i^2, $i = 1,2$.

Consider the problem of combining independent unbiased estimators of the covariance matrix. Let (12.8) be the case of interest, then (12.9) defines two independent unbiased estimators for R, the unknown covariance matrix. From (12.1) and (12.6)

$$\theta = R_2[R_1 + R_2]^{-1}\theta_1 + R_1[R_1 + R_2]^{-1}\theta_2. \tag{12.12}$$

Consider the following theorem

THEOREM 12.1 Suppose $X_1, X_2, \ldots, X_N$ $(N \geq p + 1)$ are distributed independently each according to $N(\mu,R)$. Then the distribution of

$$S = \frac{1}{N-1} \sum_{\alpha=1}^{N} (X_\alpha - \overline{X})(X_\alpha - \overline{X})^T$$

is $W(\frac{1}{N-1} R, N - 1)$, that is, the Wishart distribution with covariance matrix R and degrees of freedom $N - 1$.

It can be shown [3] that if $S = \{S_{ij}\}$ is defined in Theorem 12.1 that

$$ES = R \tag{12.13}$$

$$[S,S^T] = E\{(S_{ij} - ES_{ij})(S_{k\ell} - ES_{k\ell})\} \tag{12.14}$$

$$= \frac{1}{N_i - 1} \{\sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}\}$$

where (S,S^T) is a $(p^2 + p)x(p^2 + p)$ matrix. Defining

$$\text{(12.15)} \qquad \hat{\theta}_i = \begin{bmatrix} S_{11}^i \\ S_{22}^i \\ \vdots \\ S_{pp}^i \\ S_{12} \\ \vdots \\ S_{p-1,p} \end{bmatrix} \qquad i = 1,2$$

we note that

$$E\hat{\theta}_i = \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \vdots \\ \sigma_{pp} \\ \sigma_{12} \\ \vdots \\ \sigma_{p-1,p} \end{bmatrix} \qquad i = 1,2$$

where $R = \{\sigma_{ij}\}$. The $(\theta_i,\theta_i^T) = [S,S^T]$ as defined in (12.14). From (12.12) and (12.14)

$$\text{(12.16)} \qquad \hat{\theta} = \frac{1}{N_2 - 1} C \frac{1}{N_1 - 1} C + \frac{1}{N_2 - 1} C \quad {}^{-1}\hat{\theta}_1$$

$$+ \frac{1}{N_1 - 1} C \quad \frac{1}{N_1 - 1} C + \frac{1}{N_2 - 1} C \quad {}^{-1}\theta_2$$

where $C = \{\sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}\}$. The quantity (12.16) reduces

$$\hat{\theta} = \frac{N_1 - 1}{N_1 + N_2 - 2}\hat{\theta}_1 + \frac{N_2 - 1}{N_1 + N_2 - 2}\hat{\theta}_2 \tag{12.17}$$

a vector analog of the "pooled" estimate (12.12). Note again that the estimate $\hat{\theta}$ is independent of the covariance matrices of $\hat{\theta}_1$ and $\hat{\theta}_2$.

12.3 Recursive Estimation of the Covariance Matrix

Consider the linear model

$$y_t = H_t X_t + V_t \tag{12.18}$$

where y_t denotes a $p \times 1$ vector of observations

H_t denotes a $p \times 2$ known mapping matrix

X_t denotes an $n \times 1$ unknown state vector which we wish to estimate

V_t denotes $p \times 1$ random vector such that $EV_t = \phi$ and $EV_tV_t^T = R$ for all t. Also $EV_tV_{t+T}^T = \phi$, $T \neq 0$.

Our purpose is to estimate R from a sequence of observations ordered in time. Let

$$\{y_t;\ t = 1,2,\ldots,N\} \tag{12.19}$$

be the sequence of observations. If H_t is full rank, that is, H_t is rank $n > p$, then

(12.20) $$\hat{X}_t = (H_t^T R^{-1} H_t)^T H_t^T R_t^{-1} y_t.$$

But unfortunately in many applications the matrix R is unknown, hence must be estimated.

Another case which is of interest is the estimation of the state vector in dynamic linear filtering problems [3]

(12.21) $$\tilde{X}_t = [H_t^T R^{-1} H_t + \overline{P}_t^{-1}]^{-1} [H^T R^{-1} H_t + \overline{P}^{-1} X_t] \quad p < n$$

where X_t is an $n \times 1$ vector a priori estimate of X_t prior to taking the observations such that

$$EX_t = X_t$$

and

$$\overline{X}_t X_t^T = \overline{P}_t$$

a known covariance matrix.

Let

(12.22) $$\hat{V}_t = y_t - H_t \overline{X}_t$$

where $Ey_t \overline{X}_t^T = \phi$ and $E\overline{X}_t y_t^T = \phi$.

We note that

$$E\hat{V} = E\{y_t - H_t \overline{X}_t\} = H_t \overline{X}_t - H_t \overline{X}_t = \phi.$$

However,

$$E\hat{V}_t\hat{V}_t = E\{(y_t - H_tX_t)(y_t - H_tX_t)^T\}$$

$$= E\{y_ty_t^T - H_tX_ty_t^T - y_tX_th_t^T + H_tX_tX_t^TH_t^T\}$$

$$E\hat{V}_t\hat{V}_t^T = R + H_t\overline{P}_tH_t^T \tag{12.23}$$

An unbiased estimator can be found from (12.23) by subtracting out the bias $H_t\overline{P}_tH_t^T$, that is

$$S_t = \sum_{i=1}^{t} [\hat{V}_i\hat{V}_i^T - H_i\overline{P}_iH_i^T]/t.$$

The estimators (12.20) and (12.21) are then modified by substituting S_t for R. The analog for (12.20) is

$$\hat{X}_s = (H_t^TS_t^{-1}H_t)^{-1}H_t^TS_t^{-1}y_t \tag{12.24}$$

and the analog for (12.23) is

$$\tilde{X}_s = (H_tS_t^{-1}H_t + \overline{P}_t^{-1}) \quad [H_tS_t^{-1}y_t + \overline{P}_t^{-1}X_t]. \tag{12.25}$$

The properties of S_t, $\hat{X}_s$ and $\hat{X}_s$ have not been developed at this time.

12.4 References

[1] Graybill, F. A., An Introduction to Linear Statistical Models, Vol. 1, McGraw-Hill, 1961, pp. 409-410.

[2] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley, 1958, pp. 154-161.

[3] Tapley, B. D., and Odell, P. L., Texas Center for Research Quarterly Progress Report No. 1, June, 1964.