1980

OPTIMIZATION PROBLEMS IN STATISTICS.

KAI SANG. WONG
University of Windsor

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

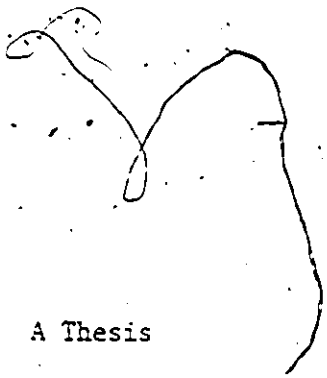
La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

OPTIMIZATION PROBLEMS IN STATISTICS

by

KAI SANG WONG



A Thesis

Submitted to the Faculty of Graduate Studies through the Department of Mathematics in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at The University of Windsor.

Windsor, Ontario

1980

740380

ABSTRACT

The theory of symbolic matrix derivatives is connected to the theory of differentials. It is shown that symbolic matrix derivatives are nothing but linear transformations of the representations of certain differentials. Representations of various differential rules are obtained and compared with those obtained by various authors. As illustrations, particular attention is given to the product rule. The theory of monotone operators is used to find the optimal solutions of various optimization problems in statistics. Some algebraic results which might be of interest by themselves are obtained to prove the main results. Optimal control models of regression experiments are presented to illustrate optimization problems with solutions on the boundary of the region of concern.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank Professor Cervin, Professor Gillen, Professor Ponnapalli, and Professor Traynor for their interest in my work.

I would like to express my sincere thanks to Professor Lemire and Professor Atkinson, who, as past and present chairmen of the Department of Mathematics, have been giving me help and encouragement during my four and a half year stay in the University.

Finally, I would like to express my deep sense of gratitude to my supervisor, Professor Chi Song Wong, who constantly gave me guidance, encouragement, help and rigorous training in mathematical statistics.

I acknowledge the financial support received from the Natural Sciences and Engineering Research Council.

TABLES OF CONTENTS

	Page
ABSTRACT	iii.
ACKNOWLEDGEMENTS	iv.
CHAPTER	
0. INTRODUCTION	1
I. REPRESENTATIONS OF DIFFERENTIALS	5
1.1 Preliminaries	5
1.2 Differentials	8
1.3 Differential Rules	9
1.4 Quadratic Differentials	12
1.5 Representations	13
1.6 Examples	30
II. MINIMA AND MAXIMA IN LINEAR MODELS AND MULTIVARIATE ANALYSIS	34
2.1 Introduction	34
2.2 Preliminaries	34
2.3 Contingency Tables	38
2.4 Maximum Likelihood Estimates of Multivariate Normal Model	43
2.5 Multivariate Regression Models	44
2.6 Multivariate Linear Hypotheses	46

TABLE OF CONTENTS (CONT'D)

		Page
2.7	Quadratic Estimates	50
2.8	Minimum Distances and Principal Components	52
2.9	Factor Analysis	55
2.10	Growth Curve Models	57
2.11	Simultaneous Equation Models	59
III.	OPTIMAL CONTROL OF A REGRESSION EXPERIMENT	62
3.1	Introduction	62
3.2	Optimal Control of a Regression Experiment	62
3.3	D-, A-, and D_s - Optimal Models	69
3.4	Reproducing Kernel Hilbert Spaces	78
REFERENCES		93
VITA AUCTORIS		

CHAPTER ZERO

Introduction

Symbolic matrix derivatives emerge as means to handle statistical and mathematical problems when the variables in a hypothesized model are many. As modern science and technology advance, the need to take more factors than one into consideration while setting up any mathematical or statistical model becomes increasingly pronounced. Multivariate analysis, originated with the paper of Hotelling (1931, see, e.g., Anderson (1958)), has become the central and one of the most important branches of statistics and data analysis. Though multivariate models are usually simple in form, the computation of test statistics and estimates is often difficult. Estimation, for instance, is usually confined to either maximum likelihood or Bayesian methods depending, partially, on one's philosophical belief. The maximum likelihood estimates are values which maximize a given likelihood function. The maximum likelihood method is more familiar to the practitioners in the field presumably because Bayesian methods are more recent and less known, and because it is employed in the classical book by Anderson (1958), or more recently, in the book by Kshirsagar (1972). The maximum likelihood method has the additional advantage that it is connected directly to hypothesis testing. A third and the oldest method, the least square method, which is prominent in the theory of linear models, is equivalent to the maximum likelihood method under the normal theory. This might be another reason for the dominance of the maximum likelihood method in multivariate analysis.

The need for matrix differentiation was pointed out in the paper by Dwyer and MacPhail (1948). In this paper, the authors define two kinds of derivatives symbolically, and apply it to the problem of least squares, canonical correlations and orthogonal regression. They also give two tables to illustrate various derivatives of real-valued functions with matrix variables or vice versa. Following the above approach, Dwyer (1967) explores further applications of matrix derivatives and works out various formulae which are useful in multivariate analysis. The connection between matrix derivatives and Jacobians was examined. Results similar to those of Deemer and Olkin (1951) (originated by P.L. Hsu) are also obtained. Later on, Tracy and Dwyer (1969) consider the derivatives of vectors with respect to vectors and use it to represent derivatives of matrix-valued functions with respect to matrices. This also leads to the second order derivatives of real-valued functions with respect to vectors or matrices. They give several applications in multivariate analysis with an attempt to justify that the critical values of the matrix-valued functions they find give the absolute minima (maxima). Following these presentations, Tracy and Singh (1972) and Singh (1972) generalize certain results to the case of partitioned matrices. In the period between the publication of these papers, the need for matrix differentiation is now well recognized and is referred to by Anderson (1958), Rao (1973) and Graybill(1969).

Econometricians also need matrix differentiation in the development of their theory. For example, Neudecker wrote three papers in the period of 1967 to 1969. In the investigation of

matrix differentiation, he uses the differential notions algebraically. In the 1969 paper, he also puts certain elements of a matrix in vector form to represent the derivatives of matrix-valued functions of matrix variables.

The paper by Vetter in 1970 is worth mentioning. In the paper, he derives a chain rule and differential rules for matrix product and Kronecker product and gives several examples to illustrate the applications of his differential rules. His work is aimed at the applications of matrix differentiation to system and control theory (see, for examples, Athans and Schweppe (1965), Athans and Tse (1967) and Athans (1967)). In 1973, McDonald and Swaminathan presented a system of matrix calculus and labelled them as McD.-S. calculus. In their paper, they give their own definitions of matrix derivatives and derive a chain rule and various product rules. Later, MacRae (1974), McDonald (1976), Swaminathan (1976) and Bentler and Lee (1975, 1978) all try to formulate and develop matrix derivatives further in this direction.

However, while the techniques of matrix derivatives are applied to various optimization problems in statistics or other disciplines such as econometrics, there is a lack of justification for the optimality of the solutions obtained by using matrix derivatives. The formulae given by various authors are long, complicated and difficult to remember. Another disadvantage with the existing methods is that there is no unity in the matrix calculus developed by various authors. Each researcher has his own basic definitions and formulae. Such a situation could lead to confusion and jeopar-

dize the development, understanding and aplications of the theory. One purpose of this dissertation is to connect the theory of matrix calculus to the familiar theory of multidimensional calculus (see, e.g., Apostal (1957) and Fleming (1977)) and linear algebra which can be treated and referred to as finite dimensional functional analysis. We shall show that the symbolic matrix derivatives mentioned above are nothing but linear transformations of the representations of certain differentials. The theory of monotone operators developed in the sixties (see, for example, Opial (1967)) will be used to find the optimal solutions of various optimization problems in statistics (Wong (to appear), Wong and Wong (1979, to appear)). We obtain some new algebraic results which are interesting in their own right. Optimal control models of regression experiments related to Chang (1979), Dorogovcev (1971) and Kiefer (1974) are presented here to illustrate optimization problems with solutions on the boundary of the region of concern. A problem raised in Chang and Wong (1979) in this connection is solved.

Chapter one will be devoted to matrix differentials and its representations. Chapter two will be devoted to the applications of differentials to the maximum likelihood theory and certain other optimization problems in statistics. Chapter three will be devoted to the solutions of certain problems of optimal control of a regression experiment.

For the sake of completeness, we include certain related results of Dr. Chi Song Wong, some of which are published and some to be published.

CHAPTER ONE

Representations of Differentials

1.1 Preliminaries

Let $L, L_1, L_2, \ldots, L_n$ be vector spaces over the real field R. Recall that the product $\prod L_i$ of L_i's is the linear space Ω of all functions (x_i) on $\{1, 2, \ldots, n\}$ such that $x_i \in L_i$, $i = 1, 2, \ldots, n$. Each L_i will be considered as a linear subspace of Ω by identifying each $x_i \in L_i$ with $f_i(x_i)$ in Ω, where f_i is the isomorphism of L_i into Ω such that $(f_i(x_i))(j) = 0$ if $j \neq i$, $f_i(x_i)(i) = x_i$. Thus Ω is the direct sum $L_1 \oplus L_2 \oplus \ldots \oplus L_n$ of $L_1, L_2, \ldots, L_n$. Let Λ be a function of Ω into L. Λ is said to be <u>multilinear</u> (<u>bilinear</u> when n=2) if Λ is coordinatewise linear on each L_i, i.e., Λ is linear in x_i when all other x_k's are fixed. $\Lambda((x_i))$ will be written as $x_1 \wedge x_2 \wedge \ldots \wedge x_n$.

Let I be a nonempty finite set. R^I will denote the family of all functions of I into R and will be equipped with the usual pointwise scalar multiplication and addition. R^I is a finite dimensional vector space over R. When $I = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\}$, R^I is the linear space M_{mxn} of all mxn matrices (over R). When $n = 1$, M_{mxn} will be denoted by R^m. In general, f in $R^{J \times K}$ is called a <u>J x K matrix</u>. $\{u_i\}_{i \in I(L)}$, $\{u_{\ell,j}\}_{j \in I(L_\ell)}$

will be bases of L, L_ℓ respectively. Let $x \in L$. Then $x = \Sigma x_i u_i$ for some unique x_i's in R. The function $[x] \equiv \{x_i\}$ on I(L) is called the linear representation of x with respect to $\{u_i\}$.

Theorem 1.1.1. [ ] above is an isomorphism of L onto $R^{I(L)}$.

Ω will be equipped with the basis $\{u_{\ell,j}\}_j \in I(L_\ell)$, $\ell = 1, 2, 3 \ldots, n$. Suppose that Λ is multilinear. Then there exists a function $[\Lambda] \equiv (a_{i,j_1,j_2,\ldots,j_n})$ of the Cartesian product $I = I(L) \times \prod_{\ell=1}^{n} I(L_\ell)$ into R such that for any u_{ℓ,j_ℓ}'s.

$$u_{1,j_1} \wedge u_{2,j_2} \wedge \ldots \wedge u_{n,j_n} = \sum_{i \in I(L)} a_{i,j_1,j_2,\ldots,j_n} u_i.$$

$[\Lambda]$ is called the linear representation of Λ (with respect to the given bases).

Theorem 1.1.2. [ ] above is an isomorphism of the family of all multilinear functions of Ω into L onto R^I.

Suppose that n = 1. Then $\Lambda \in \mathcal{L}(L_1, L)$, i.e. Λ is a linear transformation of L_1 into L. For any $f \in R^{I \times J}$, $g \in R^{J \times K}$, the (matrix)product fg of f,g is defined as an element in $R^{I \times K}$ such that each

$$(fg)((i,k)) = \sum_{j \in J} f((i,j))g((j,k)).$$

Theorem 1.1.3.

(a) $[\Lambda(x)] = [\Lambda]\,[x]$, $x \in L_1$, $\Lambda \in \mathcal{L}(L_1, L)$.

(b) For any $\Lambda_1 \in \mathcal{L}(L, L_2)$ and $\Lambda_2 \in \mathcal{L}(L_1, L)$,

$$[\Lambda_1 \circ \Lambda_2] = [\Lambda_1]\,[\Lambda_2],$$

where $\circ$ is the composition for functions.

(c) $[\ ]$ is an isomorphism of the linear space $\mathcal{L}(L_1, L)$ onto $R^{I(L) \times I(L_1)}$. Hence $R^I \times I$ is an algebra which is isomorphic to $\mathcal{L}(L, L)$.

Theorem 1.1.4. Let

$$(f, g) = \sum_{i \in I} f(i)g(i), \quad f, g \in R^I.$$

Then $(\ ,\ )$ is an inner product and R^I with $(\ ,\ )$ is a Hilbert space.

Every Hilbert space will be equipped with an orthonormal basis. The usual orthonormal basis for R^I is $\{e_i\}$:

$$e_i(j) = \delta_{ij}, \quad i, j \in I,$$

where δ_{ij}'s are the Kronecker signs. When $I = J \times J$, $(\ ,\ )$ above is called the <u>trace inner product</u> for R^I and the norm induced by $(\ ,\ )$ is called the <u>trace norm</u> for R^I.

1.2 Differentials

Let L,M be nontrivial finite dimensional Hilbert spaces (over R). Let $\{u_i\}_{i \in I(L)}$ be an orthonormal basis of L. Let C be an open set of L and f be a function of C into M. Let $x \in C$. f is said to be <u>differentiable</u> at x and has <u>differential</u> df(x) if there exists a linear transformation $df(x) \in \mathcal{L}(L,M)$ such that

$$\lim_{h \to o} \frac{f(x + h) - f(x) - df(x)(h)}{\|h\|} = 0 \quad (\in M) .$$

Here $\|\cdot\|$ is the norm induced by the inner product in L. Let $\{v_j\}_{j \in I(M)}$ be an orthonormal basis of M. Then $f(x) = \sum_{j \in I(M)} f_j(x) v_j$ for some unique $f_j(x)$,s. f(x) will be denoted by $(f_j(x))$ and f will be denoted by (f_j). $\frac{\partial f_j(x)}{\partial x_i}$ will denote

$$\lim_{t \to o} \frac{f_j(x + tu_i) - f_j(x)}{dt} \quad (\in R)$$

and is called the <u>partial derivative</u> of f_i at the point x in the direction of u_i. Suppose that df(x) exists. Then

$$(df(x)(u_i), v_j) = \frac{\partial f_j(x)}{\partial x_i}$$

and so the representation [df(x)] of df(x) is

$$[df(x)] = \left(\frac{\partial f_j(x)}{\partial x_i} \right)_{(j,i) \in I(M) \times I(L)}$$

Let A be a subset of C. We say that $f \in A^{(1)}$ if all partial derivatives $\frac{\partial f_j(x)}{\partial x_i}$ of f are continuous for all $x \in A$; $f \in A^{(2)}$ if $df \in A^{(1)}$. d(df) will be denoted by d^2f.

1.3 Differential Rules

In the following, we shall present a series of rules which are direct generalizations of the corresponding ones in real variable calculus. With the usual calculus, we could prove easily all of the following rules algebraically.

Theorem 1.3.1. (Linear rule). Let L,M be nontrivial finite dimensional Hilbert spaces. Let A be a subset of L, f,g be functions into M such that $f,g \in A^{(1)}$. Then, for every $x \in A$, $dx \in L$, $\alpha,\beta \in R$,

$$d(\alpha f + \beta g)(x)(dx) = \alpha df(x)(dx) + \beta dg(x)(dx).$$

Theorem 1.3.2. (Chain rule). Let L,M,H be nontrivial finite dimensional Hilbert spaces. Let A be a subset of L,B be a subset of M,g,f be functions into H and M respectively such that $f \in A^{(1)}$, $g \in B^{(1)}$. Then, for every $dx \in L$, $x \in A$ with $y = f(x) \in B$,

$$d(g(f(x)))(dx) = dg(y)(df(x)(dx)).$$

Theorem 1.3.3. (Rule for linear functions). Let L,M be nontrivial finite dimensional Hilbert spaces and f be a linear transformation of L into M. Then, for every x, dx ε L,

$$df(x)(dx) = f(dx).$$

Let L, L_1, L_2, . . ., L_n be montrivial finite dimensional Hilbert spaces, $\Omega = L_1 \oplus L_2 \oplus \ldots \oplus L_n$, A $\subset$ Ω, f be a function into L such that f ε $A^{(1)}$. Let $x = (x_i)$ be an element of the domain of f,

$$g_i(u) = f((x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_n)').$$

The differential of g_i at $u = x_i$ will be denoted by $\partial_{x_i} f(x)$.

Theorem 1.3.4. (Leibniz's rule). Let L, L_1, . . ., L_n be nontrivial finite dimensional Hilbert spaces. Let $A \subset \Omega \equiv \prod_{i=1}^{n} L_i$ ($= L_1 \oplus L_2 \oplus \ldots \oplus L_n$) and f be a function into L such that f ε $A^{(1)}$. Then, for every $x = (x_i)$ ε A, $dx = (dx_i)$ ε Ω,

$$df(x)\ (dx) = \sum_{i=1}^{n} \partial_{x_i} f(x)\ (dx_i).$$

The following result follows from Theorems 1.3.3 and 1.3.4.

Theorem 1.3.5. (Rule for multilinear functions). Let L, L_1, L_2, . . ., L_n be nontrivial finite dimensional Hilbert spaces. Let Λ be a multilinear function of $\Omega \equiv \prod_{i=1}^{n} L_i$ into L. Then for every (x_i), $(dx_i) \in \Omega$,

$$d(x_1 \Lambda x_2 \Lambda \ldots \Lambda x_n)((dx_i)) = \sum_{i=1}^{n} x_1 \Lambda x_2 \Lambda \ldots \Lambda x_{i-1} \Lambda dx_i \Lambda x_{i+1} \Lambda \ldots \Lambda x_n.$$

The following rule follows from Theorems 1.3.2 and 1.3.5.

Theorem 1.3.6. (Multivariate product rule). Let M, L_1, . . ., L_n be nontrivial finite dimensional Hilbert spaces. Let Λ be a multilinear function of $\Omega = \prod_{i=1}^{n} L_i$ into L. Let $A \subset M$ and f_i be functions into L_i such that $f_i \in A^{(1)}$, $i = 1, 2, \ldots, n$. Then, for every $x \in A$, $dx \in M$,

$$d(f_1(x) \Lambda \ldots \Lambda f_n(x))(dx) = \sum_{i=1}^{n} f_1(x) \Lambda f_2(x) \Lambda \ldots \Lambda f_{i-1}(x) \Lambda (df_i(x)(dx)) \Lambda f_{i+1}(x) \Lambda \ldots \Lambda f_n(x).$$

The following result is a special case of Theorem 1.3.6.

Theorem 1.3.7. (Product rule). In Theorem 1.3.6, suppose that $n = 2$. Then

$$df_1(x) \Lambda f_2(x)(dx) = df_1(x)(dx) \Lambda f_2(x) + f_1(x) \Lambda df_2(x)(dx).$$

We shall now generalize the usual Hadamard product $*$ and the Kronecker product $\otimes$ for matrices: $*$ is nothing but the pointwise product for R^I; $\otimes$ is a function of $R^{I \times J} \times R^{K \times L}$ into $R^{(I \times K) \times (J \times L)}$ such that for any $f \in R^{I \times J}$, $g \in R^{K \times L}$, each

$$(f \otimes g)(((i,j),(k,\ell))) = (f(i,j)g(k,\ell)).$$

$*$ and $\otimes$ are obviously bilinear. The products $\circledast$ and $\boxtimes$ defined in Singh (1972), Tracy and Singh (1972), the product $\odot$ defined in Khatri and Rao (1968) and the product $**$ defined in Swaminathan (1976) are all bilinear. The usual matrix product, inner product and many others are also bilinear. So the above product rule can be applied to all of them.

1.4 Quadratic Differential Forms

Let L be a nontrivial finite dimensional Hilbert space and A be an open subset of L. Let f be a real-valued function such that $f \in A^{(2)}$. Let $x \in A$. We can calculate $d^2f(x)$ from $df(x)$. However, as we shall see in Chapter Two, it is more convenient to calculate $d^2f(x)$ through the quadratic form $Q(T)$ of $T = d^2f(x)$:

$$Q(T)(dx) = (dx, T(dx)) \ , \ dx \in L.$$

Theorem 1.4.1. Let L be a nontrivial finite dimensional Hilbert space, A be an open subset of L. Let f be a real-valued function such that $f \in A^{(2)}$. Let $\{u_i\}_{i \in I(L)}$ be an orthonormal basis for L. Let $x \in A$, $dx \in L$. Then

(i) $[d^2f(x)] = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)$ and $[d^2f(x)]$ is symmetric,

where each $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ denotes $\frac{\partial}{\partial x_i}\left(\frac{\partial f(x)}{\partial x_j}\right)$.

(ii) $Q(d^2f(x))(dx) = \partial_{dx}(df(x)(dx))(dx)$.

$[d^2f(x)]$ above is called the Hessian matrix of f at x with respect to $\{u_i\}$.

1.5 Representations

Theorem 1.5.1 (Linear rule). In Theorem 1.3.1,

$$[d(\alpha f + \beta g)(x)] = \alpha[df(x)] + \beta[dg(x)],$$

i.e.,

$$[d(\alpha f + \beta g)(x)(dx)] = \alpha[df(x)][dx] + \beta[dg(x)][dx].$$

Theorem 1.5.2. (Chain rule). In Theorem 1.3.2,

$$[dg(f(x))] = [dg(y)][df(x)],$$

i.e.,

$$[dg(f(x))(dx)] = [dg(y)][df(x)][dx].$$

Theorem 1.5.3. (Rule for linear functions). In Theorem 1.3.3,

$$[df(x)] = [f],$$

i.e.,

$$[df(x)(dx)] = [f][dx].$$

Theorem 1.5.4. (Leibniz's rule). In Theorem 1.3.4,

$$[df(x)] = [[\partial_{x_i} f(x)]]$$

(a partitioned $I(L) \times \left(\prod_{i=1}^{n} I(L_i)\right)$ matrix), i.e.,

$$[df(x)(dx)] = \sum_{i=1}^{n} [\partial_{x_i} f(x)]\,[dx_i].$$

Theorem 1.5.5. (Rule for multilinear functions). In Theorem 1.3.5,

$$[d(x_1 \wedge x_2 \wedge \ldots \wedge x_n)] = [[\; x_1 \wedge x_2 \wedge \ldots \wedge x_{i-1} \wedge . \wedge x_{i+1} \wedge \ldots \wedge x_n]]$$

(a partitioned $I(L) \times \left(\prod_{i=1}^{n} I(L_i)\right)$ matrix), i.e.,

$$[d(x_1 \wedge x_2 \wedge \ldots \wedge x_n)(dx)] = \sum_{i=1}^{n} [x_1 \wedge \ldots \wedge x_{i-1} \wedge . \wedge x_{i+1} \wedge \ldots \wedge x_n][dx_i].$$

Theorem 1.5.6. (Multivariate product rule). In Theorem 1.3.6,

$$[d(f_1(x)\Lambda f_2(x)\Lambda\ldots\Lambda f_n(x))] = \sum_{i=1}^{n} [f_1(x)\Lambda\ldots\Lambda f_{i-1}(x)\Lambda.\Lambda f_{i+1}(x) \Lambda\ldots\Lambda f_n(x)][df_i(x)],$$

i.e.,

$$[d(f_1(x)\Lambda f_2(x)\Lambda\ldots\Lambda f_n(x))(dx)] = \sum_{i=1}^{n} [f_1(x)\Lambda\ldots\Lambda f_{i-1}(x)\Lambda.\Lambda f_{i+1}(x) \Lambda\ldots\Lambda f_n(x)][df_i(x)][dx].$$

Theorem 1.5.7. (Product rule). In Theorem 1.3.7,

$$[d(f_1(x)\Lambda f_2(x))] = [.\Lambda f_2(x)][df_1(x)] + [f_1(x)\Lambda.][df_2(x)],$$

i.e.,

$$[d(f_1\Lambda f_2)(x)(dx)] = [.\Lambda f_2(x)][df_1(x)][dx] + [f_1(x)\Lambda.][df_2(x)][dx].$$

Let us use the term "theory of differentials" to denote the usual theory of differentials and the related topics in linear algebra, and use the term "theory of matrix derivatives" to denote those results on matrix derivatives obtained in various papers in statistics, such as Dwyer and MacPhail (1948), Dwyer (1967), Dwyer and Tracy (1969), Neudecker (1967, 1968, 1969), Vetter (1970), MacDonald and Swaminathan (1973), MacRae (1974). Some results in the theory of differentials have been presented in section 1.1.1 - 1.5.7. Most of these results are reformulations of familiar results from undergraduate analysis (Theorem 1.3.5 is exercise 2-14 of Spivak (1965))

and are presented in such a way that all of them appear to be trivial. Our main contribution here is to connect the theory of differentials to the theory of matrix derivatives. Such a connection justifies, corrects, simplifies and generalizes the theory of matrix derivatives. As an illustration, we shall pay particular attention to Theorem 1.5.7.

Lemma 1.5.8. Let Λ be a bilinear function of $M_{m_1 x n_1} \times M_{m_2 x n_2}$ into $R^{I \times J}$. Let $[\Lambda] = (a^{(r,s)}_{(i,j),(k,\ell)}) = (a_{(r,s),((i,j),(k,\ell))})$ be the representation of Λ with respect to the usual basis. Then

(i) $a^{(r,s)}_{(i,j),(k,\ell)} = \delta_{ri}\delta_{jk}\delta_{\ell s}$, if Λ is the usual matrix product with $n_1 = m_2$.

(ii) $a^{(1,1)}_{(i,j),(k,\ell)} = \delta_{ik}\delta_{j\ell}$, if Λ is the trace inner product.

(For simplicity, we shall write $a_{(i,j),(k,\ell)}$ for $a^{(1,1)}_{(i,j),(k,\ell)}$.)

(iii) $a^{(r,s)}_{(i,j),(k,\ell)} = \delta_{rk}\delta_{\ell s}\delta_{ir}\delta_{sj}$, if Λ is the Hadamard product.

(iv) $a^{(i',k'),(j',\ell')}_{(i,j),(k,\ell)} = \delta_{ii'}\delta_{jj'}\delta_{kk'}\delta_{\ell\ell'}$, if Λ is the Kronecker product.

Proof.

(i) Let $\{e^{(i,j)}\}$, $i=1, 2,\ldots, m_1$, $j=1, 2,\ldots, n_1$, $\{f^{(k,\ell)}\}$, $k=1, 2,\ldots,n_1$, $\ell=1, 2,\ldots, n_2$ be the usual bases of $M_{m_1 x\, n_1}$ and $M_{n_1 x\, n_2}$ respectively. Let $\{g^{(r,s)}\}$, $r=1, 2,\ldots, m_1$, $s=1, 2,\ldots, n_2$ be the usual basis of $M_{m_1 x\, n_2}$. Then

$$e^{(i,j)}f^{(k,\ell)} = \sum_{r,s} a^{(r,s)}_{(i,j),(k,\ell)} g^{(r,s)}.$$

So

$$(e^{(i,j)}f^{(k,\ell)})_{\alpha\beta} = \sum_{r,s} a^{(r,s)}_{(i,j),(k,\ell)} g^{(r,s)}_{\alpha\beta}$$

$$= \sum_{r,s} a^{(r,s)}_{(i,j),(k,\ell)} \delta_{r\alpha}\delta_{s\beta}$$

$$= a^{(\alpha,\beta)}_{(i,j),(k,\ell)}.$$

On the other hand,

$$(e^{(i,j)}f^{(k,\ell)})_{\alpha\beta} = \sum_{u} e^{(i,j)}_{\alpha u} f^{(k,\ell)}_{u\beta}$$

$$= \delta_{i\alpha}\delta_{jk}\delta_{\alpha\beta}.$$

So

$$\delta_{ri}\delta_{jk}\delta_{\ell s} = a^{(r,s)}_{(i,j)(k,\ell)}.$$

(ii) Let $\{e^{(i,j)}\}$ $i=1, 2,\ldots, m$, $j=1, 2,\ldots, n$ be the usual basis of M_{mxn}. Then

$$a_{(i,j)(k,\ell)} = e^{(i,j)} e^{(k,\ell)}$$

$$= \sum_{\alpha,\beta} e^{(i,j)}_{\alpha\beta} e^{(k,\ell)}_{\alpha\beta}$$

$$= \delta_{i\alpha}\delta_{j\beta}\delta_{k\alpha}\delta_{\ell\beta}$$

$$= \delta_{ik}\delta_{j\ell}.$$

(iii) Let $\{e^{(i,j)}\}$ i=1, 2,... m, j=1, 2,..., n be the usual basis of M_{mxn}. Then

$$(e^{(i,j)} \ast e^{(k,\ell)})_{\alpha\beta} = e^{(i,j)}_{\alpha\beta} e^{(k,\ell)}_{\alpha\beta}$$

$$= \delta_{i\alpha}\delta_{j\beta}\delta_{k\alpha}\delta_{\ell\beta},$$

and

$$(e^{(i,j)} \ast e^{(k,\ell)})_{\alpha\beta} = \sum_{r,s} a^{(r,s)}_{(i,j)(k,\ell)} e^{(r,s)}_{\alpha\beta}$$

$$= a^{(r,s)}_{(i,j)(k,\ell)}\delta_{r\alpha}\delta_{s\beta}.$$

So

$$a^{(r,s)}_{(i,j)(k,\ell)} = \delta_{ir}\delta_{js}\delta_{kr}\delta_{\ell s}.$$

(iv) Let $\{e^{(i,j)}\}$ i=1, 2,..., m_1, j=1, 2,..., n_1, $\{f^{(k,\ell)}\}$ i=1, 2,..., m_2, j=1, 2,..., n_2 and $\{g^{(j'',\ell'')}_{(i'',k'')}\}$ i''=1, 2,.. m_1, k''=1, 2,... m_2, j''=1, 2,... n_1, ℓ''=1, 2,... n_2 be the usual basis of $M_{m_1 x n_1}$, $M_{m_2 x n_2}$ and $R(\{1, 2,...,m_1\} \times (\{1,2,...,m_2\}) \times (\{1, 2,...,n_1\} \times \{1, 2,..., n_2\})$ respectively. Then

$$(e^{(i,j)} \otimes f^{(f,\ell)})_{((i',k'),(j',\ell'))}$$

$$= \sum_{(i'',k''),(j'',\ell'')} a^{(i'',k''),(j'',\ell'')}_{(i,j)(k,\ell)} g^{(i'',k''),(j'',\ell'')}_{((i',k'),(j',\ell'))}$$

$$= a^{((i',k')(j',\ell'))}_{(i,j)(k,\ell)}$$

and

$$(e^{(i,j)} \otimes f^{(k,\ell)})_{((i',k')(j',\ell'))} = \delta_{ii'}\delta_{jj'}\delta_{kk'}\delta_{\ell\ell'}.$$

The required result follows. q.e.d.

Before presenting the representations of the differentials of various matrix products, we shall rewrite Theorem 1.5.7 in such a form that Lemma 1.5.8 can readily be used. Let

$$A \in R^{I_1 x I_2 x \ldots x I_k},$$

$$B \in R^{I_i}, \quad A \textcircled{i} B = \Big(\sum_{j_i \in I_i} a_{(j_1,j_2,\ldots,j_k)} b_{j_i} \Big)$$

$$\in R^{I_1 x \ldots x I_{i-1} x I_{i+1} x \ldots x I_k},$$

$$B \textcircled{i} A = A \textcircled{i} B.$$

For example, let $A = (a_{ijk})$, $i,j,k = 1,2$, with $a_{111} = 1$, $a_{112} = 2$, $a_{121} = 3$, $a_{122} = 4$, $a_{211} = 5$, $a_{212} = 10$, $a_{221} = 7$, $a_{222} = 9$, $B = (b_\ell)$ $\ell = 1, 2$ with $b_1 = 1$, $b_2 = -1$. Then $B \textcircled{1} A = (c_{jk})$ with $c_{11} = -4$,

$c_{12} = -8$, $c_{21} = -4$ and $c_{22} = 15$. Also $B \textcircled{2} A = (d_{ik})$ with $d_{11} = -2$, $d_{12} = -2$, $d_{21} = -2$, $d_{22} = 1$. $B \textcircled{1} A$ and $B \textcircled{2} A$ could be put into matrix form as follows:

$$B \textcircled{1} A = \begin{pmatrix} -4 & -8 \\ -4 & -5 \end{pmatrix}$$

and

$$B \textcircled{2} A = \begin{pmatrix} -2 & -2 \\ -2 & 1 \end{pmatrix}.$$

Lemma 1.5.9. Let Λ be a bilinear function of $R^{I_1 x I_2} \times R^{I_3 x I_4}$ into R^{IxJ}. Let $A \in R^{I_1 x I_2}$, $B \in R^{I_3 x I_4}$. Then

$$[A \Lambda .] = A \textcircled{2} [\Lambda] \quad , \quad [. \Lambda B] = [\Lambda] \textcircled{3} B.$$

Proof.

Let $\mathcal{B}_1 = \{d^{k\ell}\}$, $\mathcal{B}_2 = \{e^{rs}\}$, $\mathcal{B} = \{g^{uv}\}$ be the usual bases of $R^{I_1 x I_2}$, $R^{I_3 x I_4}$ and R^{IxJ} respectively. Write

$$[\Lambda] = (a_{(u,v),(k,\ell),(r,s)}) = (a^{(u,v)}_{(k,\ell),(r,s)}),$$

$$A = \sum_{k,\ell} a_{k\ell} d^{k\ell}, \quad B = \sum_{r,s} b_{rs} e^{rs},$$

where $a_{k\ell}$, $b_{rs} \in R$. Then

$$d^{k\ell}\wedge B = \sum_{r,s} b_{rs}(d^{k\ell}\wedge e^{rs})$$

$$= \sum_{r,s} b_{rs} \sum_{u,v} a^{(u,v)}_{(k,\ell),(r,s)} g^{uv}$$

$$= \sum_{u,v} \left(\sum_{r,s} b_{rs} a^{(u,v)}_{(k,\ell),(r,s)} \right) g^{uv}.$$

So

$$[.\wedge B] = \left(\sum_{r,s} b_{rs} a^{(u,v)}_{(k,\ell),(r,s)} \right)((u,v),(k,\ell))$$

$$= [\wedge] \textcircled{3} B.$$

Similarly,

$$A \wedge e^{rs} = \sum_{k,\ell} a_{k\ell}[d^{k\ell} \wedge e^{rs}]$$

$$= \sum_{k,\ell} a_{k\ell} \sum_{u,v} a^{(u,v)}_{(k,\ell),(r,s)} g^{uv}$$

$$= \sum_{u,v} \left(\sum_{k,\ell} a_{k\ell} a^{(u,v)}_{(k,\ell),(r,s)} \right) g^{uv}.$$

So

$$[A\wedge .] = \left(\sum_{k,\ell} a_{k\ell} a^{(u,v)}_{(k,\ell),(r,s)} \right)((u,v),(k,\ell))$$

$$= A \textcircled{2} [\wedge].$$

q.e.d.

Theorem 1.5.10. (Product rule). In Theorem 1.5.7, suppose that $L_1 = R^{I_1 x I_2}$, $L_2 = R^{I_3 x I_4}$, $L = R^{IxJ}$. Then with the usual bases,

$$[d(f_1(x)\wedge f_2(x))] = ([\wedge] \textcircled{3} f_2(x))\,[df_1(x)] + (f_1(x) \textcircled{2} [\wedge])\,[df_2(x)].$$

We close this section with a series of corollaries showing certain specializations of Theorem 1.5.10. The usual bases for R^I's will be assumed.

Corollary 1.5.11. (Inner product rule). In Theorem 1.5.10, suppose that Λ is the trace inner product. (Whence $L_1 = L_2$ and $L = R$.) Then

$$[d(f_1(x)\cdot f_2(x))] = f_1(x)[df_2(x)] + f_2(x)[df_1(x)].$$

Proof.

Let $Y \in L_1$. By Lemma 1.5.8,

$$[\Lambda] ③ Y = \left(\sum_{k,\ell} Y_{k\ell}\delta_{ik}\delta_{j\ell} \right)$$

$$= (Y_{ij})$$

$$= Y.$$

Similarly,

$$Y ② [\Lambda] = \left(\sum_{i,j} Y_{ij}\delta_{ik}\delta_{j\ell} \right)$$

$$= (Y_{k\ell})$$

$$= Y.$$

Therefore, by Theorem 1.5.10,

$$[d(f_1(x)\cdot f_2(x))] = f_1(x)[df_2(x)] + f_2(x)[df_1(x)].$$ q.e.d.

An equivalent version of corollary 1.5.11 can be found in Bentler and Lee (1978).

Corollary 1.5.12. (Matrix product rule). In Theorem 1.5.10, suppose that Λ is the matrix product on $M_{m_1 \times n_1} \times M_{n_1 \times n_2}$ Then

$$[d(f_1(x)f_2(x))] = (I_{m_1} \otimes (f_1(x))')[df_1(x)] + (f_1(x) \otimes I_{n_2})[df_2(x)].$$

Proof.

Let $Y \in M_{m_1 \times n_1}$; $Z \in M_{n_1 \times n_2}$. Then by Lemma 1.5.8,

$$Y \textcircled{3} [\Lambda] = \left(\sum_{k,\ell} Y_{k\ell}\delta_{ri}\delta_{jk}\delta_{\ell s} \right)$$

$$= (Y_{js}\delta_{j\ell})$$

$$= I_{m_1} \otimes Y',$$

$$Z \textcircled{2} [\Lambda] = \left(\sum_{i,j} Z_{ij}\delta_{ri}\delta_{jk}\delta_{\ell s} \right)$$

$$= (Z_{rk}\delta_{\ell s})$$

$$= Z \otimes I_{n_2}.$$

So by Theorem 1.5.10,

$$[d(f_1(x)f_2(x))] = (I_m \otimes (f_2(x))')[df_1(x)] + (f_1(x) \otimes I_{n_2})[df_2(x)].$$

An equivalent version of corollary 1.5.12 can be found in McDonald and Swaminathan (1973).

For any $A \in R^{I_j}$, $B \in R^{\prod_{i=1}^{n} I_i}$, the Hadamard product $A *_j B$ (or write as $B *_j A$) is defined as the element in $R^{\prod_{i=1}^{n} I_i}$ such that

$$A *_j B = (a_j b_{i_1,i_2,\ldots, i_{j-1},i_j,i_{j+1},\ldots,n}).$$

For example, let

$$I_1 = \{1,2\} \times \{1,2\},\ I_2 = \{1,2\} \times \{1,2,3\},$$

$$C = (c_{(i,j),(k,\ell)}) \in R^{I_1 \times I_2}$$

$$D = (d_{i,j}) \in R^{I_1}.$$

with

$$c_{(i,j)(k,\ell)} = ijk\ell,$$

$$d_{ij} = i+j.$$

Then

$$D *_1 C = ((i+j)(ijk\ell)) \in R^{I_1 \times I_2}$$

and can be written as the following 4 x 6 matrix by arranging the indices of the rows and columns of $D *_1 C$ in lexicographical order:

$$\begin{pmatrix} 2 \times 1 & 2 \times 2 & 2 \times 3 & 2 \times 2 & 2 \times 4 & 2 \times 6 \\ 6 \times 1 & 6 \times 2 & 6 \times 3 & 6 \times 2 & 6 \times 4 & 6 \times 6 \\ 6 \times 1 & 6 \times 2 & 6 \times 3 & 6 \times 2 & 6 \times 4 & 6 \times 6 \\ 16 \times 1 & 16 \times 2 & 16 \times 3 & 16 \times 2 & 16 \times 4 & 16 \times 6 \end{pmatrix},$$

i.e.,

$$\begin{pmatrix} 2 & 4 & 6 & 4 & 8 & 12 \\ 6 & 12 & 18 & 12 & 24 & 36 \\ 6 & 12 & 18 & 12 & 24 & 36 \\ 16 & 32 & 48 & 32 & 64 & 96 \end{pmatrix}.$$

Corollary 1.5.13. (Hadamard product rule). In Theorem 1.5.10, suppose that Λ is the Hadamard product on $M_{mxn} \times M_{mxn}$. Then

$$[d(f_1(x)*f_2(x))] = f_1(x)*_1[df_2(x)] + f_2(x) *_1[df_2(x)],$$

i.e.,

$$[d(f_1(x)*f_2(x))] = \|f_1(x)\|[df_2(x)] + \| f_2(x)\|[df_1(x)],$$

where $\|A\|$ is defined to be the mn x mn diagonal matrix whose diagonal elements $d_{ii}=a_i$, the entry of a 1 x mn vector $A = (a_i)$ at the position i. (See Bentler and Lee (1978)).

Proof.

Let $Y \in M_{mxn}$. By Lemma 1.5.7,

$$[\Lambda] \textcircled{3} Y = \left(\sum_{k,\ell} Y_{k\ell}\delta_{ik}\delta_{j\ell}\delta_{ir}\delta_{sj} \right)$$

$$= (Y_{ij}\delta_{ir}\delta_{sj})_{((r,s),(i,j))}.$$

Similarly,

$$Y \otimes [\Lambda] = (Y_{k\ell}\delta_{ir}\delta_{sj}).$$

So by Theorem 1.5.10 ,

$$[d(f_1(x)*f_2(x))] = f_1(x)*_1[df_2(x)] + f_2(x)*_1[df_1(x)]. \quad \text{q.e.d.}$$

For any $A \in R^{(I_1xI_2)x(I_3xI_4)}$, $B \in R^{I_5xI_6}$, define

$$A \otimes_1 B = (a_{(i_1,i_2),(i_3,i_4)}b_{(i_5,i_6)}) \in R^{((I_1xI_5)x(I_2xI_6))x(I_3xI_4)},$$

$$A \otimes_2 B = (a_{(i_1,i_2),(i_3,i_4)}b_{(i_5,i_6)}) \in R^{(I_1xI_2)x((I_3xI_5)x(I_4xI_6))},$$

$$B \otimes_1 A = (a_{(i_1,i_2),(i_3,i_4)}b_{(i_5,i_6)}) \in R^{((I_5xI_1)x(I_6xI_2))x(I_3xI_4)},$$

$$B \otimes_2 A = (a_{(i_1,i_2),(i_3,i_4)}b_{(i_5,i_6)}) \in R^{(I_1xI_2)x((I_5xI_3)x(I_6xI_4))}.$$

Corollary 1.5.14. (Kronecker product rule). In Theorem 1.5.10, suppose that Λ is the Kronecker product on $R^{I_1xI_2} \times R^{I_3xI_4}$. Then

$$[df_1(x) \otimes f_2(x)] = ((I_{m_1} \otimes I_{n_1}) \otimes_1 f_2(x))[df_1(x)] +$$

$$(f_1(x) \otimes_1 (I_{m_2} \otimes I_{n_2}))[df_2(x)].$$

(1.5.1)

Proof.

Let $\{e^{ij}\}$ $i=1, 2, \ldots m_1$, $j=1, 2, \ldots n_1$, $\{h^{k\ell}\}$ $k=1, 2, \ldots m_2$, $\ell=1, 2, \ldots n_2$ be the usual basis for $R^{I_1 x I_2}$ and $R^{I_3 x I_4}$ respectively. Then, from the proof of Lemma 1.3.7,

$$a^{(r,s)}_{(i,j),(k,\ell)} = (e^{ij} \otimes h^{k\ell})_{rs}.$$

Let $Y \in R^{I_1 x I_2}$, $Z \in R^{I_3 x I_4}$. Then

$$[\Lambda] \textcircled{3} Z = \left(\sum_{k,\ell} Z_k \, \delta_{ii'}, \delta_{jj'}, \delta_{kk'}, \delta_{\ell\ell'} \right)$$

$$= (Z_{k'\ell'} \delta_{ii'} \delta_{jj'})$$

$$= (I_{m_1} \otimes I_{n_1}) \otimes_1 Z.$$

Similarly,

$$Y \textcircled{2} [\Lambda] = Y \otimes_1 (I_{m_2} \otimes I_{n_2})[df_2(x)].$$

By Theorem 1.5.10 , we obtain

$$[df_1(x) \otimes f_2(x)] = ((I_{m_1} \otimes I_{n_1}) \otimes_1 f_2(x))[df_1(x)] +$$

$$(f_1(x) \otimes_1 (I_{m_2} \otimes I_{n_2}))[df_2(x)].$$

q.e.d.

We shall now compare the Kronecker product rule obtained by Bentler and Lee (1978) with corollary 1.5.14. It can be proved that the derivative of a matrix-valued function f with respect to

a matrix variable x given by Bentler and Lee (1978) is, in terms of our notations, the transpose of $[df(x)]$. The Kronecker product rule in Bentler and Lee (1978) states that, in terms of our notations in corollary 1.5.14 ,

$$[d(f_1(x) \otimes f_2(x))] = [[df_1(x)](I_{m_1 n_1} \otimes \overline{f_2(x)}) + [df_2(x)] (\overline{f_1(x)} \otimes I_{m_2 n_2})][I_{m_1} \otimes E^{n_2 m_1} \otimes I_{n_2}], \tag{1.5.2}$$

where $\bar{A}$ is the $1 \times mn$ row vector $[A_1, A_2, \ldots, A_m]$ with each A_i the i^{th} row of an $m \times n$ matrix A, E^{mr} is the $mr \times mr$ matrix such that, for $1 \leq g \leq mr$ and $1 \leq h \leq mr$, $e_{gh} = 1$, if $g = r(j-1) + k$, $h = m(k-1) + j$ $(0 < j \leq m ; 0 < k \leq r)$, and $e_{gh} = 0$ otherwise. E^{mr} is called a <u>commutation matrix</u> in Magnus and Neudecker (1979) and also appears in Tracy and Dwyer (1969) and MacRae (1974) where the notation $I_{(m,r)}$ is used. As an example, let $m = 2$, $r = 3$. Then

$$E^{2x3} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

In order to compare (1.5.1) with (1.5.2), we consider $(I_{m_1 n_1} \otimes \overline{f_2(x)})(I_m \otimes E^{n_2 m_1} \otimes I_n)$ and $(I'_{m_1} \otimes I_{n_1}) \otimes_1 f_2(x)$.

Since

$$((I_{m_1} \otimes I_{n_1}) \otimes_1 f_2(x))' = (\delta_{ij}\delta_{k\ell} f_{2\alpha\beta})((j,\ell),((i,k)(\alpha,\beta)))$$

and

$$(I_{m_1} \otimes \overline{f_2(x)})(I_{m_1} \otimes E^{n_2 m_1} \otimes I_{n_2})$$

$$= (\delta_{jj'}\delta_{\ell\ell'} f_{2\alpha'\beta'})((j,\ell)((j',\ell'),(\alpha',\beta'))$$

$$\cdot (\delta_{ij'}\delta_{\ell' k}\delta_{\alpha\alpha'}\delta_{\beta\beta'})(((j',\ell')(\alpha',\beta')),((i,k),(\alpha,\beta)))$$

$$= \Big(\sum_{j',\ell',\alpha',\beta'} \delta_{jj'}\delta_{\ell\ell'} f_{2\alpha'\beta'}\delta_{ij'}\delta_{\ell' k}\delta_{\alpha\alpha'}\delta_{\beta\beta'}\Big)((j,\ell),((i,k),(\alpha,\beta)))$$

$$= (\delta_{ij}\delta_{k\ell} f_{2\alpha\beta}).$$

$$((I_{m_1} \otimes I_{n_1}) \otimes_1 f_2(x))' = (I_{m_1} \otimes \overline{f_2(x)})(I_{m_1} \otimes E^{n_2 m_1} \otimes I_{n_2}). \tag{1.5.3}$$

Similarly,

$$(f_1(x) \otimes_1 (I_{m_2} \otimes I_{n_2}))' = (\overline{f_1(x)} \otimes I_{m_2 n_2})(I_{m_1} \otimes E^{n_2 m_1} \otimes I_{n_2}). \tag{1.5.4}$$

(1.5.3) and (1.5.4) give a shorter and rigorous proof of the result (1.5.2) of Bentler and Lee (1978). It is more important to realize that although (1.5.1) and (1.5.2) represent the very same product rule, our (1.5.1) is simpler than (1.5.2).

In this section, we have used various product rules, especially the Kronecker product rule to illustrate the simplicity of our general, rigorous and practical approach. We can carry out similar conclusions for many rules other than the Kronecker product rule.

1.6 Examples

Example 1.6.1. Let $X, dX \in M_{pxq}$, $f(X) = X'$. Then

$$df(X)(dX) = (dX)'$$

and

$$[df(X)] = (\delta_{is}\delta_{jr})_{((r,s),(i,j))}$$

Example 1.6.2. Let $X, dX \in M_{pxp}$, $f(X) = |X|$. Then

$$df(X)(dX) = tr(adjX)'dX.$$

Hence if X^{-1} exists, then

$$df(X) = |X| \, trX^{-1}dX$$

and

$$[df(X)] = |X| \, (X^{-1})'.$$

The above result can easily be proved by Leibniz's rule and the rule for linear functions. The main result in Golberg (1972) follows from Example 1.6.2 and the chain rule.

Example 1.6.3. Let $dX \in M_{qxp}$, $A \in M_{pxq}$, $B \in M_{pxr}$. Then, since $X \to AXB$ is linear,

$$d(AXB)(dX) = A(dX)B.$$

Now by corollary 1.5.12 ,

$$\begin{aligned}[d(AXB)] &= (I_q \otimes B')\,[d(AX)] \\ &= (I_q \otimes B')\,((A \otimes I_p)[dX]) \\ &= (I_q \otimes B')\,(A \otimes I_p) \\ &= A \otimes B'.\end{aligned}$$

Example 1.6.4. Let $X, dX \in M_{pxp}$ such that X^{-1} exists. Then

(i) $$d(\ln |X|\,)(dX) = \operatorname{tr} X^{-1}(dX)$$

and

$$[d\ln |X|\,] = (X^{-1})'.$$

(ii) $$dX^{-1}(dX) = -X^{-1}(dX)X^{-1}$$

and

$$[dX^{-1}] = -X^{-1} \otimes (X^{-1})'.$$

Proof.

We shall prove (ii). By Cramer's rule, it is easy to see that the inverse function is differentiable on its domain. Since $XX^{-1} = I_p$,

$$0 = d(XX^{-1})(dX)$$

$$= dX(dX)X^{-1} + X(dX^{-1}(dX))$$

$$= (dX)X^{-1} + X(dX^{-1}(dX)).$$

Hence

$$dX^{-1}(dX) = -X^{-1}(dX)X^{-1}.$$

By Corollary 1.5.12 ,

$$0 = (I_p \otimes X')[dX^{-1}] + X^{-1} \otimes I_p$$

Hence

$$[dX^{-1}] = -(I_p \otimes X')^{-1}(X^{-1} \otimes I_p)$$

$$= -(I_p \otimes (X')^{-1})(X^{-1} \otimes I_p)$$

$$= -X^{-1} \otimes (X^{-1})'. \qquad \text{q.e.d.}$$

The function f in the following example is mentioned in Bentler and Lee (1978) to challenge the existing matrix differentiation methods which had been developed up to that date.

Example 1.6.5. (Bentler and Lee (1978, p.225)). Let

$$X = (X_1, X_2, X_3, X_4, X_5), \quad \text{where} \quad X_i \in M_{m_i \times n_i}, i = 1, 2, \ldots 5,$$

$Y = (X_1 \otimes X_2)X_3(X_4 * X_5)$ such that Y^{-1} exist. Let $f(X) = trY^{-1}$. By Leibniz's rule,

$$df(X)(dX) = \sum_{i=1}^{5} \partial_{X_i} f(X)(dX_i).$$

By the rules for linear functions and inverses,

$$\partial_{X_1} f(X)(dX_1) = -\mathrm{tr}Y^{-1}\partial_{X_1} Y(dX_1)Y^{-1}$$

$$= -\mathrm{tr}Y^{-1}(dX_1 \otimes X_2)X_3(X_4 * X_5)Y^{-1}.$$

Likewise,

$$\partial_{X_2} f(X)(dX_2) = -\mathrm{tr}Y^{-1}(X_1 \otimes dX_2)X_3(X_4 * X_5)Y^{-1},$$

$$\partial_{X_3} f(X)(dX_3) = -\mathrm{tr}Y^{-1}(X_1 \otimes X_2)(dX_3)(X_4 * X_5)Y^{-1},$$

$$\partial_{X_4} f(X)(dX_4) = -\mathrm{tr}Y^{-1}(X_1 \otimes X_2)X_3(dX_4 * X_5)Y^{-1},$$

$$\partial_{X_5} f(X)(dX_5) = -\mathrm{tr}Y^{-1}(X_1 \otimes X_2)X_3(X_4 * dX_5)Y^{-1}.$$

CHAPTER TWO

Maxima And Minima in Linear Models and Multivariate Analysis

2.1 Introduction

Optimization problems in statistics can be divided up into two types: one which has the optimal solutions appeared in the interior of the region of concern and the other one which has the optimal solutions appeared on the boundary of the region of concern. For the first type, matrix differentiation is an important tool. In this chapter, we shall illustrate how matrix differentiation could be applied in various situations.

2.2 Preliminaries

In order to prove the optimality results in the following sections, we need several results which are of interest in their own right.

P_n will denote the set of all nxn positive definite matrices; S_n will denote the Hilbert space of all nxn symmetric matrices over R; SpA will denote the range of the spectrum, i.e., the set of all eigenvalues of A. It is well-known that $A \in P_n$ implies that $SpA \subset (o,\infty)$. Moreover, $Sp(AB) \subset [o,\infty)$ if $A \in P_n$, $B \in M_{nxn}$ and B is nonnegative definite.

Theorem 2.2.1. Let $A, B \in P_n$ such that $A \neq B$. Then

(i) $\quad Sp((A-B)(A^{-1}-B^{-1})) \subset (-\infty, o]$,

(ii) $\quad Sp((A-B)(A^{-1}-B^{-1})) \neq \{o\}$.

Hence,

$$tr(A-B)(A^{-1}-B^{-1}) < 0.$$

Recently, Ky Fan generalized the above result to the following:

Let A be a non-singular Hermitian matrix of order n, with p positive eigenvalues and q negative eigenvalues, $p + q = n$. Let B be a positive definite Hermitian matrix of order n, and let

$$C = (A^{-1}-B^{-1})(A-B).$$

Then all the eigenvalues of C are real, p of them non-positive, and q of them are positive. Furthermore, if exactly r of the eigenvalues of $B^{-1/2}AB^{-1/2}$ (or the similar matrix AB^{-1}) are equal to 1, then C has exactly r eigenvalues equal to zero.

Theorem 2.2.2. Let $A, B, W \in P_n$. Then

$$tr(AWA-BWB)(A^{-1}-B^{-1}) < 0.$$

Hence

$$tr(A^2-B^2)(A^{-1}-B^{-1}) < 0.$$

Proof.

Since A, B are positive definite, there exist a non-singular matrix P and a diagonal matrix D such that

$$A = PP'.$$

$$B = PDP'.$$

So

$$\begin{aligned}
&tr(AWA-BWB)(A^{-1}-B^{-1}) \\
&= tr[AW-BWBA^{-1}-AWAB^{-1}+BW) \\
&= tr\ AW(I-AB^{-1})-trBW(B^{-1}A-I) \\
&= tr\ PP'W(I-PD^{-1}P^{-1})-tr\ PDP'W(PDP^{-1}-I) \\
&= tr\ P'W(I-PD^{-1}P^{-1})P-tr\ P'W(PDP^{-1}-I)PD \\
&= tr[P'WP-P'WPD^{-1}-P'WPD^{2}+P'WPD] \\
&= tr\ P'WP(I-D^{-1}-D^{2}+D).
\end{aligned}$$

Let $E = I-D^{-1}-D^{2}+D$. Then E is diagonal. Let a_i be the entry of D at (i,i). Then the entry b_i of E at (i,i) is

$$\begin{aligned}
b_i &= 1 - \frac{1}{a_i} - a_i^2 + a_i \\
&= -\frac{1}{a_i}(a_i^3 - a_i^2 - a_i + 1) \\
&= -\frac{1}{a_i}(1 - a_i)^2(1 + a_i) \leq 0.
\end{aligned}$$

At least one of $b_i < 0$, otherwise all $a_i = 1$, i.e., $A = B$, a contradiction to $A \neq B$. Since $(c_{ij}) = P'WP$ is positive definite, all $c_{ii} > 0$, So

$$\operatorname{tr}(AWA-BWA)(A^{-1}-B^{-1}) = \sum_{i=1}^{n} c_{ii}b_i < 0. \qquad \text{q.e.d.}$$

The following result is important in justifying the differentiability of certain useful functions in statistics and is overlooked by various authors.

Theorem 2.2.3. P_n is open in S_n.

Proof.

Let $C_o = (c_{oij}) \in P_n$, $B = \{Z \in R^n : \|Z\| = 1\}$, $x \in B$. Consider $f_x : f_x(A) = x'Ax$, $A \in S_n$. Then f_x is linear on S_n and is therefore continuous.

$$|f_x(A)| = \Big|\sum_{i,j} x_i a_{ij} x_j\Big|$$

$$\leq \|A\|_\infty \sum_{j=1}^{n}\sum_{i=1}^{n} |x_i||x_j|$$

$$\leq |A|_\infty n^2.$$

Let $M = n^2$. Then $|f_x(A)| \leq M|A|_\infty$ for all $x \in B$ and $A \in S_n$. Since $x \to x'C_ox$ is continuous on R^n and B is compact, $m \equiv \min\{x'C_ox :$

$x \in B\}$ exists. Since $C_o \in P_n$, $m > 0$. Let $e = \frac{m}{2M}$, $C \in S_n$ with $\| C-C_o \|_\infty < e$, $x \in B$. Then $f_x(C) = f_x(C-C_o) + f_x(C_o) \geq -M \| C-C_o \|_\infty + m > -Me + m > 0$. So $C \in P_n$ and hence P_n is open in S_n.

q.e.d.

2.3 Contingency Tables

Consider the problem of maximum likelihood estimation and testing using likelihood ratio procedures regarding contingency tables. For simplicity, we demonstrate the case of rxk tables. The case of higher dimensions can be done similarly. The maximun likelihood function based on a sample $\{x_{ij}\}$ that forms a rxk table is

$$L(\theta) = c \prod_{i=1}^{r} \prod_{j=1}^{k} (\theta_{ij})^{x_{ij}},$$

where

$$c = \frac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{k} (x_{ij}!)},$$

$$\theta \in \Theta = \{(\theta_{ij})_{(i,j) \neq (r,k)} : \theta_{ij} > 0, \sum_{(i,j) \neq (r,k)} \theta_{ij} < 1\},$$

$$\sum_{i=1}^{r} \sum_{j=1}^{k} \theta_{ij} = 1.$$

Here x_{ij} is the number of sample values belonging to the cell I_{ij} and $n\theta_{ij}$ is the theoretical number of sample values belonging to

I_{ij}. We wish to find the maximum likelihood estimate $\hat{\theta}((x_{ij}))$ of θ, i.e., $\hat{\theta}((x_{ij})) \in \Theta$ with $L(\hat{\theta}((x_{ij}))) = \max L(\Theta)$. It can be proved that $\hat{\theta}((x_{ij}))$ exists if and only if all x_{ij}'s are positive. We shall assume that all x_{ij}'s are positive. Note that $\sum_{i,j} x_{ij}$ is equal to the sample size n. Let $f = -\ln(\frac{L}{c})$. Then $f(\hat{\theta}((x_{ij}))) = \min f(\Theta)$. Θ may be considered as an open subset of R^{rk-1}. Let $\theta \in \Theta$. Then

$$f(\theta) = -\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}\ln\theta_{ij}.$$

Let $d\theta = (d\theta_{ij}) \in R^{rk-1}$. Then by Leibniz's rule,

$$df(\theta)(d\theta) = (-\frac{x_{ij}}{\theta_{ij}} + \frac{x_{rk}}{\theta_{rk}}) \cdot (d\theta_{ij}),$$

where . is the usual inner product. Let $\theta_1 = (\theta_{1ij})$, $\theta_2 = (\theta_{2ij})$ $\in \Theta$ such that $\theta_1 \neq \theta_2$. Then

$$(df(\theta_1)-df(\theta_2))(\theta_1-\theta_2) = \sum_{(i,j)\neq(r,k)} ((-\frac{x_{ij}}{\theta_{1ij}} + \frac{x_{rk}}{\theta_{1rk}}) -$$

$$(-\frac{x_{ij}}{\theta_{2ij}} + \frac{x_{rk}}{\theta_{2rk}}))(\theta_{1ij} - \theta_{2ij})$$

$$= \sum_{(i,j)\neq(r,k)} \frac{x_{ij}(\theta_{1ij}-\theta_{2ij})^2}{\theta_{1ij}\theta_{2ij}} +$$

$$x_{rk}\left(\frac{1}{\theta_{1rk}} - \frac{1}{\theta_{2rk}}\right) \sum_{(i,j)\neq(r,k)} (\theta_{1ij}-\theta_{2ij}).$$

Since $\sum_{i=1}^{r}\sum_{j=1}^{k} \theta_{1ij} = \sum_{i=1}^{r}\sum_{j=1}^{k} \theta_{2ij} = 1$,

$$(df(\theta_1) - df(\theta_2))(\theta_1-\theta_2) = \sum_{(i,j)\neq(r,k)} \frac{x_{ij}(\theta_{1ij}-\theta_{2ij})^2}{\theta_{1ij}\theta_{2ij}} +$$

$$\frac{x_{rk}(\theta_{1rk}-\theta_{2rk})^2}{\theta_{1rk}\theta_{2rk}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{x_{ij}(\theta_{1ij}-\theta_{2ij})^2}{\theta_{1ij}\theta_{2ij}} > 0.$$

Thus, df is strictly monotone on Θ (see, e.g., Opial (1967), p.84). So f is strictly convex on Θ. Let $df(\theta) = 0$. Then

$$\frac{x_{ij}}{\theta_{ij}} = \frac{x_{rk}}{\theta_{rk}} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{k} \theta_{ij}} = n,$$

i.e., $\hat{\theta}((x_{ij})) = \left(\frac{x_{ij}}{n}\right)$.

We now consider the problem of testing the hypothesis of independence, i.e., we want to test

$$H_o : \theta_{ij} = \theta_{i.}\,\theta_{.j} \text{ for all } i=1, 2,\ldots, r,\ j=1, 2,\ldots, k.$$

against

$$H_1 : \theta_{ij} \neq \theta_{i\cdot}\theta_{\cdot j} \quad \text{for some } i, j,$$

where

$$\theta_{i\cdot} = \sum_{j=1}^{k} \theta_{ij}, \quad \theta_{\cdot j} = \sum_{i=1}^{r} \theta_{ij}.$$

Extend $\hat{\theta}$ such that $\hat{\theta}((x_{ij})) = (\frac{x_{ij}}{n})$ for all nonegative intergers with $\Sigma\, x_{ij} \leq n$, $(x_{ij}) \in R^{rk-1}$. We first assume that all x_{ij}'s are positive. Then, under H_o, the likelihood function L_1 is:

$$L_1(\alpha) = c \prod_{i=1}^{r} \prod_{j=1}^{k} (\theta_{i\cdot}\theta_{\cdot j})^{x_{ij}}$$

where

$$\alpha = ((\theta_{i\cdot}),(\theta_{\cdot j})) \in \Theta_o = \{(a,b) \in R^{r-1} \times R^{k-1} : a_i > o,\ b_i > o,$$

$$\sum_{i=1}^{r-1} a_i < 1, \quad \sum_{j=1}^{k-1} b_j < 1\},$$

$$\sum_{i=1}^{r} \theta_{i\cdot} = 1 = \sum_{j=1}^{k} \theta_{\cdot j}.$$

Note that Θ_o is open and convex in R^{r+k-2}. Let $f = -\ln(\frac{L_1}{c})$, $d\alpha = ((d\theta_{i\cdot}),(d\theta_{\cdot j})) \in \Theta_o$. Then

$$df(\alpha)(d\alpha) = \left((-\frac{x_{i\cdot}}{\theta_{i\cdot}} + \frac{x_{r\cdot}}{\theta_{r\cdot}}) , (-\frac{x_{\cdot j}}{\theta_{\cdot j}} + \frac{x_{\cdot k}}{\theta_{\cdot k}})\right) \cdot ((d\theta_{i\cdot}),(d\theta_{\cdot j})).$$

Let $\alpha_1 = ((\theta_{1i.})(\theta_{1.j}))$, $\alpha_2 = ((\theta_{2i.})(\theta_{2.j})) \in \Theta_o$ with $\alpha_1 \neq \alpha_2$. Then

$$(df(\alpha_1)-df(\alpha_2))(\alpha_1-\alpha_2) = \sum_{i=1}^{r} \frac{x_{i.}(\theta_{1i.}-\theta_{2i.})^2}{\theta_{1i.}\theta_{2i.}} + \sum_{j=1}^{k} \frac{x_{.j}(\theta_{i.j}-\theta_{2.j})^2}{\theta_{1.j}\theta_{2.j}}$$

$$> 0.$$

Therefore f is strictly convex, L is strictly concave and $(\hat{\theta}_{i.})$, $(\hat{\theta}_{.j})$ with

$$\hat{\theta}_{i.}((x_{i.})) = \frac{x_{i.}}{n}, \quad \hat{\theta}_{.j}((x_{.j})) = \frac{x_{.j}}{n}$$

are the maximum likelihood estimates of $(\theta_{i.})$, $(\theta_{.j})$ under H_o. Extend $(\hat{\theta}_{i.})$, $(\hat{\theta}_{.j})$ such that

$$(\hat{\theta}_{i.})((x_{i.})) = (\frac{x_{i.}}{n}),$$

$$(\hat{\theta}_{.j})((x_{.j})) = (\frac{x_{.j}}{n}),$$

for all $x_{i.}, x_{.j} \in \{0, 1, 2, \ldots, n\}$ such that $\Sigma x_{i.} \leq n$, $\Sigma x_{.j} \leq n$. Let

$$\lambda((x_{ij})) = \frac{L(\hat{\theta}((x_{ij})))}{L_1((\hat{\theta}_{i.})((x_{i.})),(\hat{\theta}_{.j})((x_{.j})))}$$

Then

$$\lambda((x_{ij})) = \frac{\prod_{i=1}^{r}\prod_{j=1}^{k}\left(\frac{x_{ij}}{n}\right)^{x_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{k}\left(\frac{x_{i.}}{n}\cdot\frac{x_{.j}}{n}\right)^{x_{ij}}}$$

$$= \prod_{i=1}^{r}\prod_{j=1}^{k}\left(\frac{nx_{ij}}{x_{i.}x_{.j}}\right)^{x_{ij}}.$$

It is intuitively clear that we shall reject H_o when $\lambda((x_{ij}))$ is large, say larger than a constant c. The test with the above region of rejection will be denoted by ϕ_c. The first type risk of ϕ_c is determined by c (but not vice versa). ϕ_c is often referred to as a maximum (supremum) likelihood ratio procedure (Lehman (1959), p.15).

2.4 Maximum likelihood estimates of multivariate normal model

Let $C, D \in P_n$,

$$f(C) = 1/2\ N\ln|C| - 1/2\ \mathrm{tr}CD, \quad C \in P_n.$$

In finding the maximum likelihood estimate of (μ, Σ) based on a random sample (x_α) of a normal population of mean μ and dispersion matrix Σ, it is necessary to prove that the maximum value of $f(P_n)$ is $f(ND^{-1})$ (Anderson (1958)). Let

$$g(C) = 1/2\ N\ln|C| - 1/2\ \mathrm{tr}CD, \quad C \in M_{nxn},\ |C| > 0.$$

Let h be the identity function on S_n, as a function of S_n into M_{nxn}. Then

$$f(C) = g(h(C)), \quad C \in P_n.$$

By Theorem 2.2.3 and the chain rule

$$df(C)(dC) = 1/2\ tr(NC^{-1}-D)(dC), \quad C \in P_n, \ dC \in S_n.$$

Let $C_1, C_2 \in P_n$ such that $C_1 \neq C_2$. Then by Theorem 2.2.1,

$$(df(C_1)-df(C_2))(C_1-C_2) = \frac{N}{2} tr(C_1^{-1} - C_2^{-1})(C_1-C_2)$$

$$< 0.$$

So f is strictly concave on P_n. Since $df(ND^{-1}) = 0$, $f(ND^{-1}) = \max f(P_n)$.

The role of h is important because P_n is open in S_n but has empty interior in M_{nxn}. One may compare the above proof with that of Anderson (1958, p.45-47), Smith (1978) and Watson (1964). Our proof is constructive, rigorous and simple. Note that by the result of Dykstra (1971), the maximum likelihood estimator $\hat{\Sigma}$ of Σ exists if and only if $N \geq n$.

2.5 Multivariate regression models

The likelihood function based on a sample $x_j \in N(BZ_j, \Sigma_{nxn})$ $j=1, 2, \ldots, N$ is

$$L(B,\Sigma) = (2\pi)^{-\frac{Nn}{2}} |\Sigma^{-1}|^{\frac{N}{2}} \exp\left[-1/2 \sum_{j=1}^{N} (x_j - BZ_j)' \Sigma^{-1}(x_j - BZ_j)\right],$$

where $B \in M_{nxq}$, $\Sigma \in P_n$ are unknown parameters. We want to find the maximum likelihood estimate of (B, Σ). (See, for example, Anderson (1958), Chapter 8).

Let

$$f(B,C) = \ln\left(\frac{L(B,C^{-1})}{(2\pi)^{-\frac{Nn}{2}}}\right), \quad B \in M_{nxq}, \quad C \in P_n.$$

We proceed to find $\hat{B}$ first. Fix $C \in P_n$ and let $f_C(B) = f(B,C)$. Then

$$f_C(B) = \frac{N}{2} \ln |C| - 1/2 \sum_{j=1}^{N} (x_j - BZ_j)C(x_j - BZ_j), \quad B \in M_{nxq},$$

and

$$df_C(B)(dB) = \sum_{j=1}^{N} (x_j - BZ_j)'C(dB)Z_j.$$

Let $B_1, B_2 \in M_{nxq}$. Since C is positive definite,

$$(df_C(B_1) - df_C(B_2))(B_1 - B_2) = -\sum_{j=1}^{N} Z_j'(B_1 - B_2)'C(B_1 - B_2)Z_j$$

$$\leq 0.$$

So f_C is concave on P_n. Let $df_C(B) = 0$, $dB \in M_{nxp}$. Then $df_C(B)(dB) = 0$, i.e.,

$$\sum_{j=1}^{N} (x_j - BZ_j)'C(dB)Z_j = tr(\sum_{j=1}^{N} Z_j(x_j - BZ_j)')C(dB) = 0.$$

Hence

$$\sum_{j=1}^{N} Z_j(x_j - BZ_j)'C = 0.$$

Since $C \in P_n$, $\hat{B} = (\sum_{j=1}^{N} x_j Z_j')(\sum_{j=1}^{N} Z_j Z_j')^{-1}$ is the solution for $df_C(B) = 0$. Since f_C is concave on P_n, $f_C(\hat{B}) = \max f_C(M_{nxq})$. Let

$$g(C) = f(\hat{B},C),$$

$$D = \sum_{j=1}^{N} (x_j - \hat{B}Z_j)(x_j - \hat{B}Z_j)'.$$

Then, by the result in section 2.4, $g(ND^{-1}) = \max g(P_n)$. Thus $(\hat{B}, \frac{1}{N} D)$ is the maximum likelihood estimate of (B,C).

2.6 Multivariate Linear Hypotheses

It is well known that the model discussed in section 2.5 includes certain models in regression analysis as well as analysis of variance which have various applications, especially in econometrics and psychometrics. The following related problem assumes

patterns in the dispersion matrix Σ_{nxn} of the population. It could be viewed as a general approach to the problem of analysis of variance components (see, for example, Rao (1973)). For details of the problem, one could consult Roger and Young (1978) and the references there. The log-likelihood function L based on the data Y is given by

$$L(\beta,\phi) = -\frac{N}{2}\{n \ln 2\pi - \ln|\Sigma^{-1}| + \frac{1}{N} tr(Y-H\beta)'(Y-H\beta)\Sigma^{-1}\},$$

where $\Sigma = \sum_{g=1}^{m} \phi_g G_g$, $\Sigma^{-1} = \sum_{g=1}^{m} \psi_g G_g$, G_g's are fixed, known, symmetric, linearly independent nxn matrices such that $trG_{g_1}G_{g_2} = \delta_{g_1 g_2}$ for all g_1, g_2 ; $\phi = (\phi_g)$, $\psi = (\psi_g)$ belong to an open subset G of R^m ; $m \leq \frac{n(n+1)}{2}$; $\beta \in M_{qxn}$; Σ is positive definite; $X_1, X_2, \ldots, X_n$ are independent n dimensional normally distributed random vectors with common dispersion matrix Σ and $E(Y) = H\beta$; $Y' = (X_1, \ldots, X_N)$. The problem is to find the maximum likelihood estimate $(\hat{\beta},\hat{\phi})$ of (β,ϕ). Let

$$f_\phi(\beta) = L(\beta,\phi), \quad \beta \in M_{qxn}.$$

Let $d\beta \in M_{qxn}$. Then

$$df_\phi(\beta)(d\beta) = tr\,(Y - H\beta)\,\Sigma^{-1}\,(H(d\beta))'$$

Let $\beta_1, \beta_2 \in M_{qxn}$. Since Σ^{-1} is positive definite,

$$(df_\phi(\beta_1) - df_\phi(\beta_2))\,(\beta_1-\beta_2) = -\operatorname{tr}(H(\beta_1-\beta_2))\,\Sigma^{-1}\,(H(\beta_1-\beta_2))'$$

$$\leq 0$$

Hence f_ϕ is concave on M_{qxn}. $df_\phi(\beta) = 0$ is equivalent to

$$H\,(Y-H\beta)'\,\Sigma^{-1} = 0$$

or

$$H'H\beta = H'Y$$

which is called the <u>normal equation</u> for β.

Since $H^+ = (H'H)^+ H'$,

$$\hat{\beta} = H^+Y + (I_q - H^+H)\,Z,$$

where Z is any qxn matrix and H^+ is the Moore - Penrose inverse of H. As before, let $f_{\hat{\beta}}(\psi) = L(\hat{\beta},\phi)$. Note that $\psi \to \phi$ is a one-to-one mapping and hence the maximum likelihood estimate $\hat{\phi}$ of ϕ corresponds to the maximum likelihood estimate $\hat{\psi}$ of ψ. Now, let $d\psi = (d\psi_g) \in G$. Then, by Leibniz's rule,

$$d\,\Sigma^{-1}(d\psi) = \sum_{g=1}^{m} \partial_{\psi_g}\,\Sigma^{-1}(d\psi_g)$$

$$= \sum_{g=1}^{m} G_g\,d\psi_g$$

and so

$$df_{\hat{\beta}}(\psi)\,(d\psi) = -\frac{N}{2}\left[\operatorname{tr}\Sigma\left(\sum_{g=1}^{m} G_g\, d\psi_g\right) - \frac{1}{N}\operatorname{tr} B\left(\sum_{g=1}^{m} G_g d\psi_g\right)\right],$$

where

$$B = (Y - H\hat{\beta})'\,(Y - H\hat{\beta})$$

$$= Y'\,(I - HH^{+})\,Y.$$

Let $\psi_1 = (\psi_{1g})$, $\psi_2 = (\psi_{2g}) \in G$ such that $\psi_1 \neq \psi_2$ and

$\Sigma_i = \sum_{g=1}^{m} \psi_{ig}\, G_g$, $i = 1, 2$. Then

$$(df_{\hat{\beta}}(\psi_1) - df_{\hat{\beta}}(\psi_2))\,(\psi_1 - \psi_2) = -\frac{N}{2}\left[\operatorname{tr}(\Sigma_1-\Sigma_2)\left(\sum_{g=1}^{m} G_g(\psi_{1g}-\psi_{2g})\right)\right]$$

$$= -\frac{N}{2}\left(\operatorname{tr}(\Sigma_1-\Sigma_2)\,(\Sigma_1^{-1}-\Sigma_2^{-1})\right)$$

$$< 0.$$

So $f_{\hat{\beta}}$ is strictly concave. Let $df_{\hat{\beta}}(\psi) = 0$. Then $df_{\hat{\beta}}(\psi)\,(d\psi) = 0$, for every $d\psi \in G$ and

$$\operatorname{tr}\Sigma\left(\sum_{g=1}^{m} G_g\, d\psi_g\right) = \frac{1}{N}\operatorname{tr} B\left(\sum_{g=1}^{m} G_g\, d\psi_g\right),$$

i.e.,

$$\operatorname{tr}\left(\sum_{h=1}^{m} \psi_h\, G_h\right)G_g = \frac{1}{N}\operatorname{tr} BG_g \qquad \text{for all } g = 1, 2, \ldots, m.$$

Since $\operatorname{tr} G_h G_g = \delta_{hg}$, $\frac{1}{N} \operatorname{tr} BG_g$ is the solution of $df_{\hat{\beta}}(\psi) = 0$ and, hence, is the maximum likelihood estimate $\hat{\psi} = (\hat{\psi}_g)$ of ψ.

2.7 Quadratic Estimates

Consider the regression model $Y = X\beta + \varepsilon$, where X is a known $n \times p$ matrix of rank p, ε is a normal random vector with mean 0 and dispersion matrix $\sigma^2 I$. We wish to find a quadratic estimate $Y'AY$ of σ^2 that has the smallest mean square error $E((Y'AY - \sigma^2)^2)$, $A \in S_n$. Theil and Schweitzer (1961) showed that it is equivalent to finding $A \in S_n$ such that A minimizes

$$f(A) = 2 \operatorname{tr} A^2 + (1 - \operatorname{tr} A)^2 \quad , \quad A \in S_n, AX = 0.$$

Calvert and Seber (1978) found the desired A using nearest point projections in Hilbert spaces. We shall obtain the desired A in a simple, constructive and rigorous way. Let

$$g(A) = f(A) - \operatorname{tr} N'AX,$$

where N' is a Lagrange multiplier. Let $dA \in S_n$. Then

$$dg(A)\,(dA) = 4 \operatorname{tr} A(dA) - 2(1-\operatorname{tr} A) \operatorname{tr} dA - \operatorname{tr} N'(dA)X$$

$$= \operatorname{tr} (4A - 2(1-\operatorname{tr} A)I_n - XN')dA.$$

Let A_1, $A_2 \in S_n$ such that $A_1 \neq A_2$. Since $A_1 - A_2$ is symmetric,

$$(dg(A_1) - dg(A_2))(A_1-A_2) = tr\,[4(A_1-A_2) + 2\,tr\,(A_1-A_2)]\,(A_1-A_2)$$

$$= 4tr\,(A_1-A_2)^2 + 2\cdot(tr(A_1-A_2))^2$$

$$> 0.$$

Therefore g is strictly convex on S_n. Let $dg(A) = 0$. Then

$$4A - 2(1-trA)I_n - NX' = 0. \qquad (2.7.1)$$

Multipling X on both sides of (2.7.1), we obtain, by imposing $AX=0$,

$$N = -2\,(1 - trA)\,X\,(X'X)^{-1}. \qquad (2.7.2)$$

Substituting (2.7.2) in (2.7.1), we have

$$4A - 2(1 - trA)\,[I_n - X(X'X)^{-1}\,X'] = 0.$$

So by taking the trace,

$$trA = \frac{n-p}{n-p+2}.$$

Therefore

$$\hat{A} = \frac{1}{n-p+2}\,[I_n - X(X'X)^{-1}\,X']$$

satisfies $dg(A) = 0$. Since g is convex on S_n, $g(\hat{A}) = \min g(S_n)$. Now by the theory of Lagrange multipliers (see, for example, Lehman (1959), p. 87), $f(\hat{A}) = \min f(S_n)$. Note that A is never positive definite.

2.8 Minimum Distance and Principal Components

Let $A \in P_h$. We want to find the absolute maximum (minimum) value of $f(x) = x'x$, $x \in R^h$ subject to the side condition $x'Ax = a$, where a is a fixed, positive real number. This problem is a mathematical abstraction of the principal component method in multivariate analysis (Anderson (1958, Chapter II)). In statistics, $A = \Sigma^{-1}$, where Σ is the dispersion matrix of a normal random vector. The problem is to compare the Euclidean norm with the norm $\| \ \|_\Sigma$ of the reproducing kernel Hilbert space associated with Σ. Euclidean norm is the norm of a reproducing kernel Hilbert space associated with an identity matrix. In general, let C be an $n \times n$ nonnegative definite matrix,

$$(x,y)_C = x' C^+ y, \qquad x,y \in R^h.$$

Then $(\ , \ )_C$ is a pseudo inner product, i.e., $(\ , \ )_C$ is bilinear, symmetric and $(x,x) \geq 0$ for all $x \in R^h$. $(\ , \ )_C$ is an inner product if and only if C^{-1} exists. Let $B = C^+$. Then B is nonnegative definite. Now to compare $\| \ \|_C$ with $\| \ \|_\Sigma$, we consider

$$f(x) = x' B x \qquad \text{subject to } x'Ax = a.$$

The problem raised earlier is this problem with $B = I_h$. Using the method of Lagrange multipliers, we consider

$$g(x) = f(x) + \lambda\,(a - x'Ax)$$

$$= x'Bx + \lambda\,(a - x'Ax)\,,$$

where λ is a Lagrange multiplier. Let $dx \in R^h$. Then

$$dg(x)(dx) = 2x'Bdx - 2\lambda x'A(dx)$$

$$= 2x'(B-\lambda A)(dx).$$

Let $dg(x) = 0$. Then $Bx = \lambda Ax$ or $A^{-1}Bx = \lambda x$. Suppose that we are interested in finding the absolute maximum value of f subject to $x'Ax = a$. Let $\lambda_{max} = \max Sp\,(A^{-1}B)$, e_{max} be a nonzero eigenvector of $A^{-1}B$ corresponding to λ_{max}, $c = e'_{max}\,A\,e_{max}$, $x_{max} = (\frac{a}{c})^{1/2}\,e_{max}$. Then $x'_{max}\,A\,x_{max} = a$, $B\,x_{max} = \lambda\,A\,x_{max}$ and $B - \lambda_{max}\,A$ is nonnegative definite. Let $x,y \in R^h$. Then

$$(dg(x)-dg(y))\,(x-y) = 2(x-y)'\,(B-\lambda_{max}A)\,(x-y)$$

$$\leq 0.$$

Thus g is concave on R^h and $g(x_{max}) = \max g(R^h)$. So x_{max} maximizes $x'Bx$ subject to $x'Ax = a$, $x \in R^h$. Similarly, we may use $\lambda_{min} = \min Sp\,(A^{-1}B)$ to find the absolute minimum value of f subject to $x'Ax = a$.

The above argument can be carried out for some other problems. For example, Bush and Olkin (1959), Rao (1973) and Goldberger (1964) considered the problem of minimizing $x'Ax$ subject to $B'x = u$, where $A \in P_h$ and u belongs to the column space of B'. As before, let

$$f(x) = x'Ax + \lambda'(u-B'x)$$

with λ a Lagrange multiplier. Then, for every $dx \in R^h$,

$$df(x)(dx) = 2x'A(dx) - \lambda'B'(dx)$$

$$= (2x'A - \lambda'B')(dx).$$

Let $x, y \in R^h$ such that $x \neq y$. Since A is positive definite,

$$(df(x) - df(y))\,(x-y) = 2\,(x-y)'A(x-y)$$

$$> 0.$$

So f is strictly convex on R^h. Choose $\lambda = 2(B'A^{-1}B)^{-}u$, $x = A^{-1}B(B'A^{-1}B)^{-}u$. Then $df(x) = 0$ and $B'x = u$. Hence x above minimizes $x'Ax$ subject to $B'x = u$.

In addition to principal components, we shall give another applications of the result $g(x_{max}) = \max g(R^h)$ obtained above. Using the earlier notations, let us consider

$$h(x) = \frac{x'Bx}{x'Ax}, \quad x \in R^h \setminus \{0\}.$$

Then

$$\max_{x \in R^h \setminus \{0\}} h(x) = \max_{a > 0} \max_{x'Ax=a} \frac{x'Bx}{a}$$

$$= \frac{1}{e'_{max} A e_{max}} e'_{max} B e_{max}$$

$$= \frac{e'_{max} A A^{-1} B e_{max}}{e'_{max} A e_{max}}$$

$$= \frac{e'_{max} A \lambda_{max} e_{max}}{e'_{max} A e_{max}}$$

$$= \lambda_{max}.$$

Hence for $x \in R^h \setminus \{0\}$,

$$\lambda_{min} \leq \frac{x'Bx}{x'Ax} \leq \lambda_{max}$$

and the inequalities are sharp.

2.9 Factor Analysis

In the following, we shall give an example that the likelihood function of parameters in a given model is not concave and hence the method of monotone operators fail. This pathological example shows that statisticians have to find some other method to make sure that the maximum likelihood estimates they

find are indeed the absolute maximum value of the likelihood function, especially when the estimates are obtained numerically by iteration methods.

Consider a data matrix $x = (x_{\alpha i})$ of N observations on p response variables. x is given to be an observation of a random matrix $X = (x_{\alpha i})$ with N independent rows, each having a multivariate normal distribution with the same dispersion matrix Σ of the form

$$\Sigma = \Lambda\Lambda' + \Psi,$$

where $\Psi \in P_p$ and is diagonal. This model is a familiar factor analysis model (see, for example, Lawley and Maxwell (1963)). We want to estimate Ψ. The log-likelihood function L_1, as a function of Ψ only, is

$$L_1(\Psi) = c - \frac{N}{2} \ln |\Sigma| - \frac{N}{2} \operatorname{tr} (A\Sigma^{-1}),$$

where c is a constant, A is nonnegative definite and does not depend on Σ. Let $\Psi = (\delta_{ij}\Psi_i)$. Then Ψ can be regarded as (Ψ_i) in R^p. Let $d\Psi \in R^p$. Then

$$\begin{aligned} dL_1(\Psi)(d\Psi) &= -\frac{N}{2} \operatorname{tr} \Sigma^{-1} d\Sigma(d\Psi) - \frac{N}{2} \operatorname{tr} A d\Sigma^{-1}(d\Psi) \\ &= -\frac{N}{2} \operatorname{tr} \Sigma^{-1}(d\Psi) + \frac{N}{2} \operatorname{tr} A\Sigma^{-1}(d\Psi)\Sigma^{-1} \\ &= -\frac{N}{2} \operatorname{tr} (\Sigma^{-1} - \Sigma^{-1}A\Sigma^{-1})(d\Psi). \end{aligned}$$

Let Ψ_1, Ψ_2 be two distinct values for Ψ. Then

$$\Delta \equiv (dL_1(\Psi_1) - dL_1(\Psi_2))\ (\Psi_1-\Psi_2) = -\frac{N}{2}\,[\mathrm{tr}(\Sigma_1^{-1}-\Sigma_2^{-1})(\Psi_1-\Psi_2)$$

$$- \mathrm{tr}(\Sigma_1^{-1}A\,\Sigma_1^{-1}-\Sigma_2^{-1}A\,\Sigma_2^{-1})(\Psi_1-\Psi_2)]$$

$$= -\frac{N}{2}[\mathrm{tr}(\Sigma_1^{-1}-\Sigma_2^{-1})(\Sigma_1-\Sigma_2)$$

$$-\mathrm{tr}(\Sigma_1^{-1}A\,\Sigma_1^{-1}-\Sigma_2^{-1}A\,\Sigma_2^{-1})(\Sigma_1-\Sigma_2)].$$

By Theorems 2.2.1. and 2.2.2,

$$\mathrm{tr}(\Sigma_1^{-1}-\Sigma_2^{-1})\ (\Sigma_1-\Sigma_2) < 0$$

and

$$-\mathrm{tr}\,[\Sigma_1^{-1}A\,\Sigma_1^{-1}-\Sigma_2^{-1}A\,\Sigma_2^{-1})\ (\Sigma_1-\Sigma_2)] \geq 0.$$

Since Δ is continuous in A, $\Delta > 0$ for A near 0. Thus with probability greater than 0, L is not concave.

2.10 Growth Curve Models

Consider the data matrix $x = (x_{\alpha i})$ in Section 2.9. Instead of assuming that the dispersion matrix Σ is in some structural form, we assume that

$$E(X) = A\,\xi\,K,$$

where A is an Nxh matrix of rank h and K is a gxp matrix of rank g, both being fixed matrices with $h \leq N$ and $g \leq p$. ξ is the unknown hxg parameter matrix we wish to estimate. We also assume that Σ is known. This is the familiar growth curve model considered by Potthoff and Roy (1964), (See also Joreskog (1970)). The log-likelihood function, as a function of ξ, is

$$L(\xi) = c - \frac{N}{2} \ln |\Sigma| - \frac{N}{2} \operatorname{tr} A \Sigma^{-1},$$

where c is a constant and

$$A = \frac{1}{N} (x - A \xi K)' (x - A \xi K).$$

Let $d\xi \in M_{hxg}$. Then

$$dL(\xi)(d\xi) = - \frac{N}{2} \operatorname{tr} dA(d\xi)\Sigma^{-1}$$

$$= N \operatorname{tr} (x - A\xi K)' A (d\xi)K\Sigma^{-1}.$$

Since Σ is positive definite,

$$Q(d^2L(\xi))(d\xi) = - N \operatorname{tr} (A(d\xi)K)' A (d\xi)K\Sigma^{-1}$$

$$\leq 0.$$

Thus L is concave on M_{hxg}. Let $dL(\xi) = 0$. Then

$$K\Sigma^{-1} (x - A\xi K)' A = 0$$

or

$$A'A\xi K\Sigma^{-1}K' = A'x\Sigma^{-1}K'.$$

Hence $(A'A)^{-1} Ax\Sigma^{-1}K'(K'\Sigma^{-1}K)^{-1}$ is the solution of $dL(\xi) = 0$ and is, therefore, the maximum likelihood estimate $\hat{\xi}$ of ξ.

2.11 Simultaneous Equation Models

Let $U = (u_{ij})$ be an Nxt matrix of normally distributed random variables u_{ij} with $E(U) = 0$ and $E(\frac{1}{t}UU') = \Sigma$. Consider the model

$$BY + \Gamma Z = U,$$

where B and Γ are NxN and NxΛ matrices of parameters. Assume that Σ is known. The problem is to obtain the maximum likelihood estimators of B and Γ. This is the familiar problem of full information maximum likelihood estimation of the structural parameters of a simultaneous linear structural equation model considered by Fisk (1967, Chapter 4), Neudecker (1967), Koopmans (1950) and Tracy and Singh (1972). The log-likelihood function of (B,Γ) is given by

$$L(B,\Gamma) = c + t \ln |B| + \frac{t}{2} \ln |\Sigma^{-1}| - \frac{t}{2} \operatorname{tr}\Sigma^{-1}AMA',$$

where

$$M = \frac{1}{t} \begin{bmatrix} Y \\ Z \end{bmatrix} (Y', Z'),$$

$$A = (B, \Gamma) \in M_{Nx(N+\Lambda)},$$

c is a constant and Σ, B are positive definite. Note that Neudecker (1967) and Tracy and Singh (1972) do not assume that B is positive

definite and they do not prove that the optimal solution they found are indeed the maximum likelihood estimate of B.

Let $dA \in M_{Nx(N+\Lambda)}$, $dA \neq 0$, $dB = P(dA)$. Since projection P of A to B is linear,

$$dL(A)(dA) = t\ trB^{-1}(dB)' - t\ tr\Sigma^{-1}AM(dA)'$$

Since B, M and Σ are positive definite,

$$Q(d^2L(A))(dA) = -t\ tr\ B^{-1}dBB^{-1}(dB)' - t\ tr\ \Sigma^{-1}(dA)M(dA)'$$

$$< 0.$$

Thus L is strictly convex on $M_{Nx(N+\Lambda)}$. Let $dL(A) = 0$. Then (and only then)

$$(B^{-1},0) = \Sigma^{-1}AM. \qquad (2.11.1)$$

The solution of (2.11.1) is, therefore, the maximum likelihood estimate of (B,Γ).

For simplicity, let us use the above example to demonstrate certain advantages of our approach. One reason that our approach is so simple is because the functions $A \to L(A)$, $dA \to dL(A)(dA)$, $dA \to Q(dL^2(A))(dA)$ we are dealing with are real-valued.

Neudecker shows, in terms of our notations, up to one-to-one linear transformations,

$$[d^2L(A)] = -t \begin{pmatrix} [(B')^{-1},0] \otimes B_1^{-1} \\ \\ [(B')^{-1},0] \otimes B_p^{-1} \end{pmatrix} + t'MA'\Sigma^{-1}AM \otimes \Sigma^{-1}$$

$$+ t \begin{pmatrix} \Sigma^{-1}AM \otimes \{MA'\Sigma^{-1}\}_1 \\ \\ \Sigma^{-1}AM \otimes \{MA'\Sigma^{-1}\}_q \end{pmatrix} - t\, M \otimes \Sigma^{-1}.$$

Tracy and Singh shows that, in terms of our notations, up to one-to-one transformations,

$$[d^2L(A)] = -t \begin{pmatrix} B^{-1} \otimes (B')^{-1} & 0 \\ 0 & 0 \end{pmatrix} + t\,(\Sigma^{-1} \boxtimes MA'\Sigma^{-1}AM)$$

$$+ t\,(MA'\Sigma^{-1} \boxtimes \Sigma^{-1}AM) - t\,(\Sigma^{-1} \boxtimes M)$$

Here C_i is the i^{th} row of C and $A \boxtimes B$ is defined in Tracy and Singh (1972). One can see from the above results that their approaches are by no means simple. Also, they use the usual theory of calculus without connecting it to their own theories of matrix calculus. Moreover, we show that L is strictly concave. So (2.11.1) cannot have more than one solution. Thus one can use the conventional numerical methods, such as the method of Fletcher and Powell (1963), to solve (2.11.1).

CHAPTER III

Optimal Control of a Regression Experiment

3.1 Introduction

We have seen in the preceding chapter that using the first and second order differentials, many optimization problems in statistics can be solved practically and rigorously. However, for optimal problems of the second type, often, some other methods are needed. Linear programming techniques may be used to optimize a linear function with linear inequality constraints. Variational techniques may be used to optimize a functional with linear or nonlinear inequality constraints. In this chapter, we shall use a linear regression model to illustrate certain methods of solving the optimal problems of the second type.

3.2 Optimal Control of a Regression Experiment

Consider the regression model m(f):

$$Y(t) = f(t)\,\theta + \varepsilon(t) \quad , \quad t \in T, \qquad (3.2.1)$$

where T is a nonempty topological space, $Y(t) = (y_1(t),\ldots,y_n(t))'$, $f(t) = (f_{ij}(t)) \in M_{nxk}$, $\varepsilon(t) = (\varepsilon_1(t),\ldots,\varepsilon_n(t))'$, $\theta = (\theta_1,\ldots,\theta_k)' \in R^k$. $\{\varepsilon(t)\}_{t\in T}$ is a vector stochastic process on some probability

space $(\Omega,\mathcal{A},P)$ such that each $E(\varepsilon(t)) = 0 \in R^n$, each $r(i,s,j,t) = E(\varepsilon_i(s)\varepsilon_j(t))$ is known and each $r(i,\cdot,j,\cdot)$ is continuous on TxT. We shall be interested in the control problem of choosing f in some chosen set X to minimize a certain functional $\Phi(\cdot)$ of the dispersion matrix $\Sigma(f)$ of the least square estimator $\hat{\theta}$ of θ. Chang and Wong (to appear) consider this problem for the case n=1 and Dorogovcev (1971) considers this problem for the case n=1 and k=2. Related problems also appear in Chang (1979), Chang and Wong (to appear) and Mehra (1974). When T is a singleton, the above problem is familiar in optimal design theory. See, for example, Federov (1972) and Kiefer (1974).

In practice, n represents the number of observations. It turns out that the general case ($n \geq 1$) can be reduced to the case n=1 so that the optimality results in Chang and Wong (to appear) and Chan and Wong (to appear) can be applied. For simplicity, we shall assume that T = [a,b], where $a,b \in R$ and $a<b$. We can easily generalize the obtained results to more general settings. Let W be an m-dimensional linear subspace of $\mathcal{L}^2(T)$ and $\{\eta_1,\eta_2,\ldots,\eta_m\}$ be an orthonormal basis in W. We shall assume that each η_i is continuous on T, and in (3.2.1), $f = (f_1,\ldots,f_k)$, where each $f_j = (f_{ij})$, $f_{ij} \in W$. Thus $f_{ij} = \sum_{\alpha} f_{(\alpha,i),j}\,\eta_\alpha$ for some unique $f_{(\alpha,i),j}$'s. Unless otherwise stated, we shall keep all notations. Note that all $f_j \in H \equiv W^n$.

Theorem 3.2.1. $\{f_j\}$ is independent in H if and only if $F = \{f_{(\alpha,i),j}\}_j$ is independent in M_{mxn}.

We shall assume that $\{f_j\}$ is independent in H. F above will be treated as an element in $R^{(\{1,2,\ldots,m\}x\{1,2,\ldots n\})x\{1,2,\ldots,k\}}$. Let g, h ε H, e.

$$(g,h) = \sum_{i=1}^{n} (g_i,h_i).$$

Then H with (,) is a Hilbert space. $\|\ \|$ will denote the norm induced by the above inner product (,). (,) is the usual product inner product for H. X will denote $\{(f_{ij})$: all $f_{ij} \varepsilon W$, f_j's are independent in H, $\|f_j\| \leq 1$, $j = 1,2,\ldots,k\}$, $1 > 0$. For simplicity, we shall take $1 = 1$.

Let $\omega \varepsilon \Omega$. Then the least square estimate of θ, denoted by $\hat{\theta}(\omega)$, is defined to be the θ which minimizes

$$\| Y(\cdot)(\omega) - F\theta \|^2 \quad , \quad \theta \varepsilon R^k. \qquad (3.2.2)$$

M(f) will denote F'F, where $F' = \{c_{j,(\alpha,i)}\}$ with each $c_{j,(\alpha,i)} = f_{(\alpha,i),j}$. Thus $M(f) \varepsilon M_{kxk}$. Each $U=(t_{ij}) \varepsilon R^{IxJ}$ can be considered as a linear transformation [U] of R^J into R^I such that $[U](x_j)=$

$U(x_j)(=(\Sigma t_{ij}x_j)\varepsilon\ R^I)$, $(x_j)\varepsilon\ R^J$. Thus the theory of linear transformations can be applied to R^{IxJ}. Now $r(M(f)) = r(F) = k$. So $M(f)^{-1}$ exists.

Theorem 3.2.2.

$$\hat{\theta}(\omega) = M(\cdot)^{-1}((Y(\cdot)(\omega),f_j)), \quad \omega \varepsilon \Omega . \tag{3.2.3}$$

Proof. Let $\omega \varepsilon \Omega$. Let $j=1,2,\ldots,k$, $Y = Y(\cdot)(\omega)$. Then

$$(Y - f\ \hat{\theta}(\omega),f_j) = 0 .$$

i.e.,

$$(Y,f_j) = (f\ \hat{\theta}(\omega),f_j)$$

$$= \sum_{i,\ell} (f_{i\ell},f_{ij})\ \hat{\theta}_\ell(\omega).$$

Since

$$(f_{ij},f_{i\ell}) = \int_T f_{i\ell}(t)f_{ij}(t)dt$$

$$= \sum_\alpha f_{(\alpha,i),\ell} f_{(\alpha,i),j},$$

$$(F'F)\hat{\theta}(\omega) = ((Y,f_j)).$$

Therefore

$$\hat{\theta}(\omega) = (F'F)^{-1}((Y,f_j))$$

$$= M(\cdot)^{-1}((Y,f_j)). \qquad \text{q.e.d.}$$

$\Sigma(f)$ will denote the dispersion matrix of $\hat{\theta}$.

Theorem 3.2.3.

$$\Sigma(f) = M(f)^{-1}(F'RF)M(f)^{-1},$$

where

$$R = (r_{(\alpha,i),(\beta,j)}) = (\int_T\int_T r(i,s,j,t)\eta_\alpha(s)\eta_\beta(t)dsdt) \in R^{LxL}$$

with

$$L = \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}.$$

Proof.

$$\Sigma(f) = M(f)^{-1}\Sigma_\Xi M(f)^{-1},$$

where Σ_Ξ is the dispersion matrix of Ξ with $\Xi(\omega) = ((Y(\cdot)(\omega),f_j))$.

Since each $E(\varepsilon_i(t)) = 0$, by Fubini's theorem, $E((\varepsilon(\cdot),f_j)) = 0$. Thus by Fubini's theorem,

$$\Sigma_Z = (E((\varepsilon(\cdot),f_u)(\varepsilon(\cdot),f_v)))$$

$$= (\sum_{i,\ell} E((\varepsilon_i(\cdot),f_{iu})(\varepsilon_\ell(\cdot),f_{\ell v})))$$

$$= (\sum_{i,\ell} E(\int_T\int_T \varepsilon_i(s)\varepsilon_\ell(t)f_{iu}(s)f_{\ell v}(t)dsdt))$$

$$= (\sum_{i,\ell,\alpha,\beta}\int_T\int_T E(\varepsilon_i(s)\varepsilon_\ell(t))f_{\alpha ui}f_{\beta v\ell}\eta_\alpha(s)\eta_\beta(t)dsdt)$$

$$= (\sum_{i,\ell,\alpha,\beta} f_{\alpha ui}(\int_T\int_T r(i,s,\ell,t)\eta_\alpha(s)\eta_\beta(t)dsdt)f_{\beta v\ell})$$

$$= F'RF.$$

Therefore

$$\Sigma(f) = M(f)^{-1}(F'RF)M(F)^{-1}. \qquad \text{q.e.d.}$$

With the above estimator $\hat{\theta}$ of θ, our objective is to choose an f in X such that for a certain function ϕ, $\phi(\Sigma(f)) = \min_{g\varepsilon X} \phi(\Sigma(g))$. The following are three important ϕ's in optimum design theory. These and together with some other criterions can be found in Federov (1972). Let ϕ be a convex functional on P_k. Then $m(f)$ is called a ϕ-optimal model if $\phi(\Sigma(f)) = \min_{g\varepsilon X} \phi(\Sigma(g))$. A ϕ-optimal model $m(f)$ is said to be

(i) D-Optimal model if

$$\phi(C) = |C|^{-1} \quad , \quad C \varepsilon P_k \quad ; \qquad (3.2.4)$$

(ii) A-Optimal if

$$\phi(C) = \text{tr } C^{-1} \quad ; \qquad (3.2.5)$$

(iii) D_s-Optimal model if

$$\phi(C) = |C_s|^{-1} \quad , \quad C \varepsilon P_k, \qquad (3.2.6)$$

where C_s is obtained from C by deleting the last $(k-s)$ rows and columns. ϕ_1, ϕ_2, ϕ_3 will denote respectively ϕ in (i) (ii) and (iii).

Lemma 3.2.3. Let h be a function of P_k into R such that $h \varepsilon (P_k)^{(2)}$ ($P_k \subset S_k$). Let

$$h_\ell(A) = \exp(\ell(A) + h(A)), \quad A \varepsilon P_k,$$

$\ell \in S_k^{\#}$, the conjugate space of S_k. Then the following two conditions are equivalent:

(i) h is (strictly) convex.

(ii) h_ℓ is (strictly) convex for all $\ell \in S^{\#}$.

Theorem 3.2.4. ϕ_1, ϕ_2 are strictly convex.

Proof. Let $A \in P_k$, $dA \in S_k$. Then

$$d\phi_1(A)(dA) = - \operatorname{tr} A^{-1}(dA)A^{-1}$$

$$= - \operatorname{tr} A^{-2}dA.$$

So for any distinct $A, B \in P_k$, by Theorem 2.2.2,

$$(d\phi_1(A) - d\phi_1(B))(A-B) = - \operatorname{tr}(A^{-2}-B^{-2})(A-B)$$

$$> 0.$$

Hence ϕ_1 is strictly convex on P_k. Let

$$\phi(A) = -\ln |A| \quad , \quad A \in P_k.$$

Let $A \in P_k$, $dA \in S_k$. Then

$$d\phi(A)(dA) = - \operatorname{tr} A^{-1}(dA).$$

So for any distinct $A, B \in P_k$, by Theorem 2.2.1,

$$(d\phi(A) - d\phi(B))(A-B) = - \operatorname{tr}(A^{-1}-B^{-1})(A-B) > 0.$$

Therefore Φ is strictly convex on P_k. By Lemma 3.2.3 with $\hat{\ell}=0$, Φ_2 is strictly convex. q.e.d.

Since $C \to C_s$ is linear, by Theorem 3.2.4, Φ_3 is convex.

3.3 D-, A- and D_s-Optimal models

To find the D-, A- and D_s-optimal models m(f), one has to solve for each i=1,2,3, the optimization problem $\Phi_i(f)=\max_{g\in X} \Phi_i(g)$. Using results in Chang and Wong (to appear) and Chan and Wong (to appear), the above problem can be solved.

Let H_1 be an mnxk matrix obtained by re-arranging the "rows" $f_{(\alpha,i)}$ of the "matrix" F in any preassigned order. Let R_1 be the mnxmn matrix obtained by re-arranging the "rows" and "columns" of R in the corresponding way. Since multiplication of matrices does not depend on these re-arrangements,

$$\Sigma(f) = (H_1'H_1)^{-1}H_1'R_1H_1(H_1'H_1)^{-1}.$$

Since R_1 is positive definite, there exist an orthogonal matrix P such that

$$PR_1P' = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_{mn} \end{pmatrix} \equiv D,$$

where $0<d_1 \leq d_2 \leq \ldots \leq d_{mn}$. It is important to stress here that if one changes the above arrangement of rows and columns, then R_1 will be replaced by a similar matrix R_2 whence R_1 but not D will be changed. Let $G = PH_1$. Then

$$\Sigma(f) = (G'G)^{-1}G'DG(G'G)^{-1}.$$

Moreover, the columns G_j of G has norm less than or equal to 1:

$$\|G_\beta\|^2 = \sum_{\alpha,j} f^2_{(\alpha,j),\beta}$$

$$= \sum_j \|f_{j\beta}\|^2$$

$$= \|f_\beta\|^2$$

$$\leq 1.$$

Thus the results in Chang and Wong (to appear) and Chan and Wong (to appear) for m(f) with n=1, m replaced by mn can be applied. We shall assume that D is positive definite. The following results follow trivially from Theorem 3 in Chan and Wong (to appear).

Theorem 3.3.1. Let

$$c = \frac{\sum_{j=1}^{k} d_j^{1/2}}{k},$$

$$D_1 = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_k \end{pmatrix} ,$$

$$G = \frac{1}{c^{1/2}} \begin{pmatrix} D_1^{1/4} \\ 0 \end{pmatrix} Q' ,$$

Where Q is an orthogonal matrix such that $Q\, D_1^{1/2} Q'$ has equal diagonal elements. Then m(f) is A-optimal and

$$\Sigma(f) = c\, Q\, D_1^{1/2} Q'.$$

Of course, in order to find an optimal model m(f), one must first obtain F from G and then obtain f from F.

Theorem 3.3.2. The following conditions are equivalent:

a) m(f) is D-optimal.

b) $G = Q\begin{pmatrix} U \\ 0 \end{pmatrix}$,

for some orthogonal matrix Q and sxs diagonal matrix U whose diagonal elements all belong to {-1,1}.

Theorem 3.3.3. The following conditions are equivalent:

a) m(f) is D_s-optimal.

b) $G = Q\begin{pmatrix} U & 0 \\ 0 & V \\ 0 & 0 \end{pmatrix}$,

where U is an sxs diagonal matrix such that each diagonal element of U is -1 or 1, V is a (k-s)x(k-s) nonsingular matrix such that each column has Euclidean norm equal or less than 1.

For illustration, we shall give an example. Earlier notations will be kept. Suppose that for m(f), n=2=m=k, a=0, b=1, all $\varepsilon_1(s)$, $\varepsilon_2(t)$ are independent normal random variables of mean 0 and

$$r(i,s,j,t) = r(s,t)\delta_{ij},$$

where for s,t in (0,1),

$$r(s,t) = (1-s)\ (1-t)\ \min\ \{\frac{s}{1-s}\ ;\ \frac{t}{1-t}\},$$

$$r(1,t) = r(s,1) = r(1,1) = 0.$$

Since $\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}\ (2t-1)$, $\frac{1}{\sqrt{2}} - \frac{\sqrt{3}}{\sqrt{2}}\ (2t-1)$ are orthonormal in $\mathcal{L}^2[0,1]$, we may take

$$\eta_1 = \frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}\ (2t-1),$$

$$\eta_2 = \frac{1}{\sqrt{2}} - \frac{\sqrt{3}}{\sqrt{2}}\ (2t-1).$$

We shall use Theorem 3.3.1 to find an A-optimal model m(f). Now

$$r_{(1,1),(1,1)} = \int_0^1 \int_0^1 (\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}\ (2t-1))(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}\ (2s-1))$$

$$(1-s)(1-t)\ \min\ \{\frac{s}{1-s},\ \frac{t}{1-t}\}\ dsdt$$

$$= \int_0^1 \int_s^1 \left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2t-1)\right)\left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2s-1)\right)(1-s)t\,dt\,ds$$

$$+ \int_0^1 \int_0^s \left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2t-1)\right)\left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2s-1)\right)(1-t)s\,ds\,dt$$

$$= \frac{1}{10} .$$

$$r_{(1,1),(2,1)} = \int_0^1 \int_0^1 \left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2s-1)\right)\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{3}}{\sqrt{2}}(2t-1)\right)$$

$$(1-s)(1-t) \min\left\{\frac{s}{1-s}, \frac{t}{1-t}\right\} ds\,dt$$

$$= \int_0^1 \int_s^1 \left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2s-1)\right)\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{3}}{\sqrt{2}}(2t-1)\right)(1-s)t\,dt\,ds$$

$$+ \int_0^1 \int_0^s \left(\frac{1}{\sqrt{2}} + \frac{\sqrt{3}}{\sqrt{2}}(2t-1)\right)\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{3}}{\sqrt{2}}(2s-1)\right)(1-t)s\,ds\,dt$$

$$= \frac{-1}{15} .$$

Similarly,

$$r_{(2,1),(2,1)} = r_{(2,2),(2,2)} = r_{(1,2),(1,2)} = \frac{-1}{10} ,$$

$$r_{(2,1),(1,1)} = r_{(2,2),(1,2)} = r_{(1,2),(2,2)} = \frac{-1}{15} ,$$

$$r_{(\alpha,i),(\beta,j)} = 0 \qquad \text{elsewhere.}$$

We shall use the order:

$$(1,1) \longrightarrow 1,\ (1,2) \longrightarrow 3,\ (2,1) \longrightarrow 2,\ (2,2) \longrightarrow 4.$$

R is then reduced to

$$R_1 = \begin{pmatrix} \frac{1}{10} & -\frac{1}{15} & 0 & 0 \\ -\frac{1}{15} & \frac{1}{10} & 0 & 0 \\ 0 & 0 & \frac{1}{10} & -\frac{1}{15} \\ 0 & 0 & -\frac{1}{15} & \frac{1}{10} \end{pmatrix}$$

It is easy to diagonalize

$$\begin{pmatrix} \frac{1}{10} & -\frac{1}{15} \\ -\frac{1}{15} & \frac{1}{10} \end{pmatrix} \quad \text{to} \quad \begin{pmatrix} \frac{1}{30} & 0 \\ 0 & \frac{1}{6} \end{pmatrix}$$

by

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

So R_1 can be diagonalized to

$$\begin{pmatrix} \frac{1}{30} & & & 0 \\ & \frac{1}{6} & & \\ & & \frac{1}{30} & \\ 0 & & & \frac{1}{6} \end{pmatrix}$$

by

$$A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Let P be the matrix obtained from A by interchanging the second and third columns. Then

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

and

$$D \equiv P'R_1P = \begin{pmatrix} \frac{1}{30} & 0 & 0 & 0 \\ 0 & \frac{1}{30} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & \frac{1}{6} \end{pmatrix}$$

Thus $d_1 = d_2 = \frac{1}{30}$, $d_3 = d_4 = \frac{1}{6}$, $c = \frac{(d_1^{\frac{1}{2}} + d_2^{\frac{1}{2}})}{2} = \frac{1}{\sqrt{30}}$,

$$D_1 = \frac{1}{30} I_2.$$

Let Q be a 2x2 orthogonal matrix. Then $Q D_1^{\frac{1}{2}} Q'$ has equal diagonal elements. In fact, $Q D_1^{\frac{1}{2}} Q' = \frac{1}{30} I_2$.

So

$$G = C^{-\frac{1}{2}} \begin{pmatrix} D_1^{\frac{1}{2}} \\ 0 \end{pmatrix} Q'$$

$$= \begin{pmatrix} I_2 \\ 0 \end{pmatrix} Q' .$$

$$H_1 = P'G$$

$$= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \end{pmatrix} Q'$$

For illustration, let $Q = I_2$. Then

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \end{pmatrix}$$

Using the earlier order, we obtain $F = (f_{(\alpha,i),j})$ with

$$f_{(1,1),1} = \frac{1}{\sqrt{2}} = f_{(1,1),2}$$

$$f_{(1,2),1} = \frac{1}{\sqrt{2}} = -f_{(1,2),2}$$

$$f_{(\alpha,i),j} = 0 \text{ elsewhere.}$$

Let $t \in [0,1]$. Then

$$f_{11}(t) = f_{(1,1),1}\,\eta_1(t) + f_{(2,1),1}\eta_2(t)$$

$$= \frac{1}{\sqrt{2}} \eta_1(t),$$

i.e.,

$$f_{11}(t) = \tfrac{1}{2}\,((1-\sqrt{3}) + 2\sqrt{3}\,t).$$

Similarly

$$f_{12}(t) = f_{11}(t) = f_{21}(t),$$

$$f_{12}(t) = -\frac{1}{\sqrt{2}} \eta_1(t).$$

Thus

$$f(t) = \begin{pmatrix} \tfrac{1}{2}\,((1-\sqrt{3}) + 2\sqrt{3}\,t) & \tfrac{1}{2}\,((1-\sqrt{3}) + 2\sqrt{3}\,t) \\ \tfrac{1}{2}\,((1-\sqrt{3}) + 2\sqrt{3}\,t) & -\tfrac{1}{2}\,((1-\sqrt{3}) + 2\sqrt{3}\,t) \end{pmatrix},$$

$t\in[0,1]$. With this A-optimal model $m(f)$,

$$\Sigma_{\hat{\theta}} = \Sigma(f)$$

$$= c\,QA^{\frac{1}{2}}\,Q'$$

$$= c\,D_1^{\frac{1}{2}}$$

$$= \frac{1}{30}\,I_2.$$

i.e. $\hat{\theta}_1$, $\hat{\theta}_2$ are independent random variables of variance $\frac{1}{30}$.

D-optimal and D_s-optimal models $m(f)$ can be found numerically by using Theorems 3.3.2 and 3.3.3.

3.4. Reproducing Kernel Hilbert spaces

The notion of least squares in section 3.3 does not depend on the covariance function R. There appears to have no sufficient reason to use the Euclidean norm to measure a random distance. In this section, we shall use a distance associated with R.

Instead of assuming that D is positive definite, we shall assume that $R(i,s,j,t) = R_i(s,t)\delta_{ij}$, where all R_i are continuous. So for $i \neq j$, all $y_i(s)$; $y_j(t)$ are independent. Chang (1979) deals with the case n=1. For completeness, let us define the notion of positive definiteness mentioned above. Let K be a continuous real-valued function on TxT. Note that all R_i are examples of K. Let $f \in \mathcal{L}(T)$.

$$\hat{K}(f)(t) = \int_a^b f(s)K(s,t)ds, \qquad t \in [a,b].$$

By the continuity of K and the compactness of [a,b], $\hat{K}(f)$ is continuous on [a,b]. So $\hat{K}$ is a linear function of $\mathcal{L}(T)$ into $\mathcal{L}^\infty(T)$. Since

$$\mathcal{L}^\infty(T) \subset \mathcal{L}^2(T) \subset \mathcal{L}(T),$$

$\hat{K}$, restricted to $\mathcal{L}^2(T)$, is a linear function of $\mathcal{L}^2(T)$ into itself. We shall assume that $\hat{K}$ is restricted to $\mathcal{L}^2(T)$. For simplicity, let $f \in \mathcal{L}^2(T)$, $s \in T$. Since $K(s, \cdot) \in \mathcal{L}^2(T)$, by Schwartz's inequality,

$$|\hat{K}(f)(s)| \leq \| K(s, \cdot) \| \, \| f \|.$$

Thus

$$\| \hat{K}(f) \| \leq \| K \| \, \| f \|,$$

where $\| K \|$ is the norm of K in $\mathcal{L}^2(TxT)$. Let $B(\mathcal{L}^2(T))$ be the algebra of all continuous linear functions of $\mathcal{L}^2(T)$ into $\mathcal{L}^2(T)$ with the operator norm $\| \quad \|$. Then $\hat{K} \in B(\mathcal{L}^2(T))$ and as it was shown above,

$$\| \hat{K} \| \leq \| K \| .$$

Since Λ is one-to-one, we may identify $\hat{K}$ with K and write K for $\hat{K}$. Since $K \in B(\mathcal{L}^2(T))$, the notions and theory of operators in Hilbert spaces can be applied, e.g., K is said to be <u>positive (nonnegative) definite</u> if K is self adjoint, i.e. K=K', the adjoint of K, and for any nonzero f in the Hilbert space $\mathcal{L}^2(T)$, $(Kf,f) > 0$ (≥ 0). The adjoint K' of K is defined as $K' \in B(\mathcal{L}^2(T))$ such that

$$(K'f,g) = (f,Kg) \qquad , \qquad f,g \in \mathcal{L}^2(T).$$

Thus K=K' if and only if K is symmetric. So all R_i are self adjoint. We shall now let i=1,2,...,n, $K=R_i$ and prove that K is nonnegative definite. Let $f \in \mathcal{L}^2(T)$. Then

$$(Kf,f) = \int_a^b (\int_a^b f(s)K(s,t)ds)f(t)dt$$

$$= \int_a^b \int_a^b f(s)f(t)\, E\, (\varepsilon_i(s)\varepsilon_i(t))dsdt$$

$$= E\, (\int_a^b \int_a^b f(s)f(t)\varepsilon_i(s)\varepsilon_i(t)dsdt)$$

$$= E\, ((\int_a^b f(s)\varepsilon_i(s)ds)^2)$$

$$\geq 0.$$

So K is nonnegative definite. For convenience, we shall assume that all R_i are positive definite. This assumption is not serious and amounts to: For any $i=1,2,\ldots,n$ and any $A \in \mathcal{A}$ with $P(A) = 1$, $\mathcal{L}^2(T)$ is the closed linear span of $\{\varepsilon(.)(\omega) : \omega \in A\}$. It can be shown by the theory of Wiener processes that r in the example of section 3.3 is nonnegative definite.

The background of the above discussions can be found in Riesz and Nagy (1955), especially the Chapters IV and VI, respectively, on integral equations and completely continuous symmetric transformations of Hilbert space.

Now by Mercer's theorem, for each i, there exist an orthogonal basis $\{\phi_{iu}\}$ for $\mathcal{L}^2(T)$ and a sequence $\{\lambda_{iu}\}$ of positive real numbers such that

$$R_i(s,t) = \sum_{u=1}^{\infty} \lambda_{iu}\, \phi_{iu}(s)\, \phi_{iu}(t) \qquad (3.4.1)$$

in $\mathcal{L}^2(T)$ and $\{\sum_{u=1}^{v} \lambda_{iu}\, \phi_{iu}(s)\, \phi_{iu}(t)\}$ converges uniformly to R_i on TxT. Here $\{\lambda_{iu}\}$ is the spectrum of R_i, (3.4.1) is the spectral representation of R_i and each ϕ_{iu} is an eigenvector of R_i corresponding to λ_{iu}. Let

$$H(R_i) = \{f \in \mathcal{L}^2(T) : \sum_{u=1}^{\infty} \frac{1}{\lambda_{iu}}(f,\phi_{iu})^2 < \infty\},$$

where (,) is the usual inner product for $\mathcal{L}^2(T)$. For g, h ε $H(R_i)$, we define $g_{iu} = (g,\phi_{iu})$ $(h_{iu} = (h, \phi_{iu}))$ and

$$(g,h)_{R_i} = \sum_{u=1}^{\infty} \frac{g_{iu}h_{iu}}{\lambda_{iu}} .$$

Then $H(R_i)$ is a Hilbert space and will be referred to as the reproducing kernel Hilbert space (RKHS) induced by R_i. Let $R = (R_i)$ and, as in section 3.2, let H(R) be the product of $H(R_i)$'s equipped with the product inner product $(,)_R$, i.e.,

$$(g,h)_R = \sum_{i=1}^{n} (g_i,h_i)_{R_i} \quad , \quad g_i,h_i \varepsilon H(R_i),$$

$$g = (g_i),\ h = (h_i).$$

Then H(R) is a Hilbert space and is called the reproducing kernel Hilbert space (RKHS) induced by R. $\| \ \|_R$ will denote the norm induced by $(,)_R$.

$$\|g\|_R^2 = (g,g)_R \quad , \quad g \varepsilon H. \qquad (3.4.3)$$

We assume that f_j's in m(f) are continuous on T and are independent in H(R). Let $\omega \varepsilon \Omega$. ξ in R^k, denoted by $\tilde{\theta}(\omega)$, is the Gauss - Markov estimator of θ based on the observation $Y(.)(\omega)$ if

$$\|Y(.)(\omega) - f\,\xi\|_R^2 = \min_{\theta \varepsilon R^k} \|Y(.)(\omega) - f\,\theta\|_R^2$$

Since $f_1, f_2, \ldots, f_k$ are independent in $H(R)$ and $\{f\theta : \theta \in R^k\}$ is the finite dimensional linear space S spanned by $f_1, f_2, \ldots, f_k$, $\tilde{\theta}(\omega)$ uniquely exists and with $Y = Y(.)(\omega)$,

$$(Y - f\tilde{\theta}(\omega), f_j)_R = 0 \quad , \quad j = 1,2,\ldots,k.$$

Thus

$$(Y,f_j)_R = (f\tilde{\theta}(\omega), f_j)_R$$

$$= \sum_{\ell} \tilde{\theta}_\ell(\omega) (f_\ell, f_j)_R .$$

Hence

$$(\sum_{\ell=1}^{k} \tilde{\theta}_\ell(\omega) (f_\ell, f_j)_R) = ((Y,f_j)_R).$$

Let

$$M(f) = ((f_\ell, f_j)_R) .$$

Then $M(f)^{-1}$ exists and

$$M(f)\tilde{\theta}(\omega) = ((Y,f_j)_R) .$$

Therefore

$$\tilde{\theta}(\omega) = M(f)^{-1} ((Y,f_j)_R) . \qquad (3.4.4)$$

Let W be a random vector of $(\Omega,\mathcal{A},P)$ into R^k. W is said to be a continuous linear estimator of θ if there exist linear continuous functionals $L_1, L_2, \ldots, L_k$ on H(R) such that

$$W(\omega) = (L_j(Y(.)(\omega))) \quad , \quad \omega \in \Omega.$$

Theorem 3.4.1. $\tilde{\theta}(\omega)$ is the best unbiased linear continuous estimator (BLUE) of θ.

Proof. It suffices to prove that $\tilde{\theta}(\omega)$ is unbiased for θ and that $\tilde{\theta}(\omega)$ has minimum variance among all unbiased linear continuous estimator of θ. Let $\omega \in \Omega$, and W be a continuous linear estimator of $\bar{\theta}$. Then by Riesz's representation theorem,

$$W(\omega) = ((Y(.)(\omega),g_j)_R),$$

for some $g_1, g_2, \ldots, g_k$ in H(R). Write $g_j = (g_{ij})$. Then

$$E(W) = (\int_\Omega \sum_{i=1}^{n} (Y_i(.)(\omega),g_{ij})_{R_i} dP(\omega))$$

$$= (\sum_{i=1}^{n} \int_\Omega (Y_i(.)(\omega),g_{ij})_{R_i} dP(\omega))$$

$$= (\sum_{i=1}^{n} \int_\Omega \sum_{u=1}^{\infty} \frac{g_{iju}}{\lambda_{iu}} \int_a^b Y_i(t)(\omega)\phi_{iu}(t)dtdP(\omega))$$

$$= (\sum_{i=1}^{n} \sum_{u=1}^{\infty} \frac{g_{iju}}{\lambda_{iu}} \int_a^b \sum_{\ell=1}^{k} \varepsilon_{i\ell}(t)\theta_\ell\phi_{iu}(t)dt)$$

$$= \left(\sum_{\ell=1}^{k} \theta_\ell \sum_{i=1}^{n} \sum_{u=1}^{\infty} \frac{g_{iju}}{\lambda_{iu}} f_{i\ell u}\right)$$

$$= \left(\sum_{\ell=1}^{k} \theta_\ell \sum_{i=1}^{n} (g_{ij}, f_{i\ell})_{R_i}\right)$$

$$= \left(\sum_{\ell=1}^{k} \theta_\ell \, (g_j, f_\ell)_R\right).$$

So

$$E(W) = ((f_\ell, g_j)_R)\theta. \tag{3.4.5}$$

Now by (3.4.5),

$$E(\tilde{\theta}(\omega)) = E(M(f)^{-1}((Y(.), f_j)_R))$$

$$= M(f)^{-1}((f_\ell, f_j)_R)\,\theta$$

$$= \theta,$$

i.e., $\tilde{\theta}(\omega)$ is unbiased for θ. Let Σ_W be the dispersion matrix of W. Then

$$\Sigma_W = E((\varepsilon(.)(\ ), g_j)_R, (\varepsilon(.)(\ ), g_k)_R)$$

$$= \left(\sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{u=1}^{\infty} \sum_{u'=1}^{\infty} \frac{1}{\lambda_{iu}\lambda_{iu'}} g_{iju} g_{i'ku'} \int_a^b \int_a^b \right.$$

$$\left. E(\varepsilon_i(s)\varepsilon_{i'}(t))\phi_{iu}(s)\phi_{i'u'}(t)\,ds\,dt\right)$$

$$= \left(\sum_{i=1}^{n} \sum_{u=1}^{\infty} \sum_{u'=1}^{\infty} \frac{1}{\lambda_{iu}\lambda_{iu'}} g_{iju} g_{iku'} \int_a^b \int_a^b R_i(s,t)\right.$$

$$\phi_{iu}(s)\phi_{iu'}(t)dsdt)$$

$$= \Big(\sum_{i=1}^{n} \sum_{u=1}^{\infty} \sum_{u'=1}^{\infty} \frac{1}{\lambda_{iu}\lambda_{iu'}} g_{iju}g_{iku'} \sum_{u''=1}^{\infty} \frac{1}{\lambda_{iu''}} \delta_{uu''}\delta_{u'u''}\Big)$$

$$= \Big(\sum_{i=1}^{n} \sum_{u=1}^{\infty} \frac{1}{\lambda_{iu}} g_{iju}g_{iku'}\Big).$$

So

$$\Sigma_W = ((g_j, g_k)_R). \tag{3.4.6}$$

By (3.4.6),

$$\Sigma(f) = M(f)^{-1} \Sigma_{(Y(.),f_j)_R} (M(f)^{-1})'$$

$$= M(f)^{-1} M(f) M(f)^{-1}.$$

So

$$\Sigma(f) = M(f)^{-1}. \tag{3.4.7}$$

By (3.4.5), W is unbiased for θ if and only if

$$((g_j, f_\ell)_R) = I_k. \tag{3.4.8}$$

It can be proved by (3.4.6) - (3.4.8) that if W is unbiased for θ, then $\Sigma_W - \Sigma_{\tilde{\theta}(\omega)}$ is nonnegative definite. Hence $\tilde{\theta}(\omega)$ has minimum variance among all unbiased continuous estimators of θ.

q.e.d.

To obtain D-, D_s- and A-optimal models m(f), we need to specify the family X of all eligible f. Let $L \in (0,\infty)$ and let X be the family of $f = (f_{ij})$ in (3.2.1) such that $\|f_j\|_R \le L$, $j = 1,2,\ldots,k$.

Theorem 3.4.2. (a) $\min_{g\in X} |\Sigma(g)| = \frac{1}{L^{2k}}$.

(b) m(f) is a D-optimal model if and only if $f_1, f_2, \ldots, f_k$ are orthogonal in H(R) with norm L.

Proof. (a) Since

$$|\Sigma(g)| = |M(g)|^{-1}, \qquad g \in X,$$

it suffices to prove that

$$\max_{g\in X} |M(g)| = L^{2k}.$$

Let $g \in X$ and let $\{\lambda_j\}$ be the spectrum of M(g). Then

$$\sum_{j=1}^{k} \lambda_j = \operatorname{tr} M(g) = \sum_{j=1}^{k} (f_j,f_j)_R$$

$$= \sum_{j=1}^{k} \|f_j\|_R^2$$

$$\le \sum_{j=1}^{k} L^2$$

$$= k L^2. \qquad (3.4.9)$$

So, by the familiar arithematic mean - geometric mean inequality,

$$|M(g)| = \prod_{j=1}^{k} \lambda_j$$

$$\leq \left(\sum_{j=1}^{k} \frac{\lambda_j}{k} \right)^k$$

$$\leq \left(\frac{1}{k} \cdot k L^2 \right)^k$$

$$= L^{2k}.$$

Now suppose that $f_1, f_2, \ldots, f_k$ are orthogonal in H(R) with norm L. Then

$$|M(f)| = |((f_i, f_j)_R)|$$

$$= |L^2 I_k|$$

$$= L^{2k},$$

proving (a).

(b) From (a), it suffices to prove that $\min_{g \in X} |M(g)| = |M(f)| = L^{2k}$ implies that $f_1, f_2, \ldots, f_k$ are orthogonal in H(R) with norm L. Let $\{\lambda_j\}$ be the spectrum of M(f). Since $\{\lambda_j\}$ minimizes $\prod_{i=1}^{p} \lambda_j$ with

all $\lambda_j > 0$

and

$$\sum_{j=1}^{k} \lambda_j \leq k L^2, \tag{3.4.10}$$

$\lambda_1 = \lambda_2 = \ldots = \lambda_p = L^2$. Since M(f) is positive and has identical eignevalues, $\lambda_j = L^2$, $M(f) = L^2 I_k$. Thus f_1, f_2, ..., f_k are orthogonal in H(R) and have norm L. q.e.d.

Note here that even for n=1, the above result is more general than Theorem 1 (i) in Chang (1979). Also our proof is different from the one given by Chang who uses Hadamard's inequality to prove (a).

Theorem 3.4.3.

(a) $\min_{g \in X} \operatorname{tr} \Sigma(g) = \frac{k}{L^2}$.

(b) m(f) is an A-optimal model if and only if f_1, f_2, ..., f_k are orthogonal in H(R) with norm L.

Proof. (a) Let $g \in X$. Then

$$\operatorname{tr} \Sigma(g) = \operatorname{tr} M(g)^{-1}.$$

Let $\{\lambda_j\}$ be the spectrum of M(g). Then

$$\operatorname{tr} \Sigma(g) = \sum_{i=1}^{k} \frac{1}{\lambda_i}. \tag{3.4.11}$$

To minimize tr $\Sigma(g)$ is equivalent to minimize (3.4.11) subject to (3.4.10), which, again, is equivalent to minimize (3.4.11) subject to

$$\text{all } \lambda_i > 0 \quad , \quad \sum_{i=1}^{k} \lambda_i = k\,L^2. \tag{3.5.12}$$

Let $\lambda = (\lambda_i)$ and

$$\phi(\lambda) = \sum_{i=1}^{k} \frac{1}{\lambda_i} - \mu\Big(\sum_{i=1}^{k} \lambda_i - k\,L^2\Big),$$

where μ is a Lagrange multiplier. Let $d\lambda = (d\lambda_i) \in R^k$. Then

$$d\phi(\lambda)(d\lambda) = \Big(-\frac{1}{\lambda_i^2} - \mu\Big)\, d\lambda.$$

Let $\alpha = (\alpha_i)$, $\beta = (\beta_i) \in R^k$ with all $\alpha_i, \beta_i > 0$. Then

$$(d\phi(\alpha)-d\phi(\beta))(\alpha-\beta) = \Big(\Big(\frac{1}{\beta_i^2} - \frac{1}{\alpha_i^2}\Big)\Big)\big((\alpha_i-\beta_i)\big)$$

$$= \sum_{i=1}^{k} \frac{(\alpha_i-\beta_i)^2\,(\alpha_i+\beta_i)}{\alpha_i^2\beta_i^2}$$

$$\geq 0.$$

So ϕ is convex on the open convex set B of all λ in R^k with each $\lambda_i > 0$. Let $d\phi(\lambda) = 0$. Then

$$\lambda_i = (-\mu)^{-\frac{1}{2}} \quad , \quad i = 1,2,\ldots,k.$$

Choose μ so that (3.5.12) holds. Then

$$\lambda_i = L^2 \quad , \quad i = 1,2,\ldots,k, \quad \mu = -\frac{1}{L^4} ,$$

and so λ_0 with each $\lambda_{0i} = L^2$ gives the minimum value of $\phi(B)$. By the theory of Lagrange's multipliers, λ_0 minimizes (3.4.11) subject to (3.5.12) and

$$\text{tr } \Sigma(f) \geq \sum_{i=1} \frac{1}{\lambda_{0i}} = \frac{k}{L^2} .$$

Now choose f such that $f_1, f_2, \ldots, f_k$ are orthogonal in H(R) with norm L. Then $f \in X$ and $\Sigma(f) = \frac{1}{L^2} I_k$. Thus

$$\text{tr } \Sigma(f) = \frac{k}{L^2} ,$$

proving (a).

(b) By (a), it suffices to prove that tr $\Sigma(f) = \frac{k}{L^2}$ implies that $f_1, f_2, \ldots, f_k$ are orthogonal in H(R) with norm L. From the proof of (a), the spectrum $\{\lambda_{0j}\}$ of M(f) minimizes (3.5.11) subject to (3.5.12). So

$$\lambda_{01} = \lambda_{02} = \ldots = \lambda_{0k} = L^2,$$

i.e., $M(f) = L^2 I_k$. Hence $f_1, f_2, \ldots, f_k$ are orthogonal in H(R) with norm L. q.e.d.

Again, even for n=1, Theorem 3.4.3 is more general than Theorem 1 (ii) in Chang (1979). Also, our proof is different from the one given by Chang who uses the Gram - Schmidt process to prove (a). Combining Theorem 3.4.2 and Theorem 3.4.3 we have the following nice result.

Theorem 3.4.4. Let $f \in X$. Then $m(f)$ is D-optimal if and only if it is A-optimal.

From the view point of section 3.3, where D-optimal models are not equivalent to A-optimal models, Theorem 3.4.4 is a very strong and, yet, a desirable result. One then wonders which estimator and optimal model should be used. The Gauss-Markov estimator is the BLUE of θ. However, for finding this estimator, one needs the spectral representation of all r_i, which are often difficult to obtain. Also, we seldom know all r_i precisely and if we do not know all r_i precisely, then a repeated use of r_i's will probably increase the uncertainty of the chosen model and the chosen estimator. Since the optimal models $m(f)$ obtained in this section depend on a certain RKHS, practically, the design f may be difficult to control (or, say, construct). On the other hand, it is relatively easy to control the optimal models in section 3.3 and compute the underlying least square estimates. Also, in terms of motivations and conclusions, A-optimal models $m(f)$ in section 3.3 tend to ignore larger eigenvalues of R. Indeed, with the assumption of this section, if $n \geq k$, then only the smallest eigenvalue of R

contributes to the optimal models. On the other hand, the A-optimal models m(f) in this section are obtained by more or less treating all eigenvalues of r_i's as equal. To conclude, like many other statistical models, it is up to the workers to observe the reality and then decide which optimal model to use.

REFERENCES

ANDERSON, T.W. (1958). Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York.

APOSTAL, T.M. (1957). Mathematical Analysis. Addison Wesley, Mass.

ATHANS, M. (1967). The Matrix Minimum Principle. Information and Control, 2, 592-606.

ATHANS, M. and SCHWEPPE, F.C. (1965). Gradient Matrices and Matrix Calculation. M.I.T. Lincoln Lab. Lexington, Mass. Technical Report.

ATHANS, M. and TSE, E. (1967). A Direct Derivation of the Optimal Linear Filter Using the Maximum Principle. IEEE Trans. Automatic Control, AC-12, 690-698.

BENTLER, P.M. and LEE, Sik-Yum (1975). Some Extension of Matrix Calculus. General Systems, 20, 145-50.

BENTLER, P.M. and LEE, Sik-Yum (1978). Matrix Derivatives with Chain Rules and Rules for Simple Hadamard and Kronecker Products. J. Math. Psychology, 17, 255-262.

BUSH, K.A. and OLKIN, I. (1959). Extrema of Quadratic Forms with Applications to Statistics. Biometrika, 46, 483-486.

CHAN, N.N. and WONG, Chi Song (To Appear). Existance of an A-Optimal Model for a Regression Experiment. J. Math. Analysis and Its Appl.

CHANG, D.S. (1979). Design of Optimal Control for a Regression Problem. Ann. Statist., 7, 1078-1085.

CHANG, D.S. and WONG, Chi Song (To Appear). Corrections to "Design of Optimal Control for a Regression Problem." Ann. Statist.

CHANG, D.S. and WONG, Chi Song (To Appear). A General Approach to Optimal Control of a Regression Experiment. J. Multivariate Analysis.

CHANG, D.S. and WONG, Chi Song (1979). Optimal Control of a Regression Experiment. Seminar on Functional Analysis, Institute of Mathematics, National Tsing Hua University, Taiwan, (Editor: Chi Song Wong).

CALVERT, B. and SEBER, G.A. F. (1978). Minimization of Functions of a Positive Semidefinite Matrix A Subject to AX = 0. J. Multivariate Analysis, 8, 173-180.

DEEMER, W.L. and OLKIN, I. (1951). The Jacobians of Certain Matrix Transformations Useful in Multivariate Analysis. Biometrika, 38, 345-67.

DOROGOVCEV, A. Ja. (1971). Problems of Optimal Control of a Regression Experiment. Selected Transl. in Math. Statist. and Probability, 10, 35-41.

DWYER, P.S. and MacPHAIL, M.S. (1948). Symbolic Matrix Derivatives. Ann. Math. Statist., 19, 517-34.

DWYER, P.S. (1967). Some Applications of Matrix Derivatives in Multivariate Analysis. J. Amer. Statist. Assoc., 62, 607-625.

DYKSTRA, R.L. (1970). Establishing the Positive Definiteness of the Sample Covariance Matrix. Ann. Math. Statist. 41, 5153-54.

FEDEROV, V.V. (1972). Theory of Optimal Experiments. Academic Press, New York.

FISK, P.R. (1967). Statistically Dependent Equations. Charles Griffin and Company. London.

FLEMING, W.H. (1977). Functions of Several Variables, Second Edition. Addison-Wesley, Reading.

FLETCHER, R. and POWELL, M.J.D. (1963). A Rapidly Convergent Descent Method for Minimization. Computer Journal, 6, 163-168.

GOLBERG, M.A. (1972). The Derivative of a Determinant. Amer. Math. Monthly, 79, 1124-1126.

GOLDBERGER, A.S. (1964). Econometric Theory. John Wiley and Sons Inc., New York.

GRAYBILL, F.A. (1969). Introduction to Matrices With Applications in Statistics. Wadsworth, California.

HOTELLING, H. (1931). The Generalization of Student's Ratio. Ann. Math. Statist., 2; 360-378.

*JENNRICH, R.I. (1973). Standard Error for Obliquely Rotated Factor Loadings. Psychometrika, 38, 593-603.

JORESKOG, K.G. (1970). A General Method for Analysis of Covariance Structures. Biometrika, 57, 239-251.

KIEFER, J. (1974). General Equivalence Theory for Optimal Designs. Ann. Statist., 2, 849-79.

KHATRI, C.G. and RAO, C.R. (1968). Solutions to Some Functional Equations and Their Applications to Characterization of Probability Distributions. Sankhya, Series A, 30, 167-180.

KOOPMANS, T.C. (1950). Statistical Inference in Dynamic Economic Models. Cowles Commission for Research in Economics, Monograph No. 10. John Wiley, New York.

KSHIRSAGAR, A.M. (1972). Multivariate Analysis. Mercel Dekker, New York.

LEHMAN, E.L. (1959). Testing Statistical Hypothesis. John Wiley and Son Inc., New York.

LAWLEY, D.N. and MAXWELL, A.E. (1963). Factor Analysis as a Statistical Method. Butterworths, London.

MacRAE, E. (1974). Matrix Derivatives With an Application to an Adaptive Linear Decision Problem. Ann. Statist., 2, 337-46.

MAGNUS, J.R. and NEUDECKER, H. (1979). The Commutation Matrix: Some Properties and Applications. Ann. Statist., 7, 381-394.

McDONALD, R.P. and SWAMINATHAN, H. (1973). A Simple Matrix Calculus With Applications to Multivariate Analysis. General System, 18, 37-54.

McDONALD, R.P. (1976). The McDonald-Swaminathan Calculus: Clarifications, Extensions and Illustrations. General System, 21, 87-94.

MEHRA, R.K. (1974). Optimal Input Signals for Parameter Estimation in Dynamic System -- Survey and New Results. IEEE Trans. Automatic Control, AC-19, 753-68.

*NEL, D.G. (1978). A Review of Matrix Differentiation in Statistics. Technical Report No. 38, Dept. of Math. Statist., University of Orange Free State, South Africa.

NEUDECKER, H. (1967). On Matrix Procedures for Optimizing Differentiable Scalar Functions of Matrices. Statist. Neerlandica, 21, 101-107.

NEUDECKER, H. (1968). The Kronecker Matrix Product and Some of its Applications in Econometrics. Statist. Neerlandica, 22, 69-82.

NEUDECKER, H. (1969). Some Theorems on Matrix Differentiation With Special Reference to Kronecker Matrix Products, J. Amer. Statist. Assoc., 64, 953-63.

OLKIN, I. (1953). Note on "The Jocabians of Certain Matrix Transformations Useful in Multivariate Analysis." Biometrika, 40, 43-46.

OPIAL, Z. (1967). Lecture Notes on Nonexpansive and Monotone Mapping in Banach Spaces. Center for Dynamical Systems, Brown University, Providence.

POTTHOFF, R.F. and ROY, S.N. (1964). A Generalized Multivariate Analysis of Variance Model Useful Especially for the Growth Curve Problems. Biometrika, 5, 313-16.

RAO, C.R. (1973). Linear Statistical Inference and its Applications, Second Edition. John Wiley and Sons Inc., New York.

RIESZ, F. and SZ-NAGY, B. (1955, 3rd Edition). Functional Analysis. (Translated by Boron, L,F.), Frederick Ungar Publishing Co.

ROGER, G.S. and YOUNG, D.L. (1978). On Testing a Multivariate Linear Hypothesis When the Covariance Matrix and its Inverse Have the Same Pattern. J. Amer. Statist. Assoc., 73, 203-7.

SINGH, R.P. (1972). Some Generalizations in Matrix Differentiation With Applications in Multivariate Analysis. Doctoral Dissertaton. University of Windsor.

SMITH, D.W. (1978). A Simplified Approach to the Maximum Likelihood Estimation of the Covariance Matrix. Amer. Statistician, 32, 28-29.

SPIVAK, M. (1965). Calculus of Manifolds. Benjaman, New York.

SWAMINATHAN, H. (1976). Matrix Calculus for Functions of Partitioned Matrices. General System, 21, 95-99.

THEIL, H. and SCHWEITZER, A. (1961). The Best Quadratic Estimator of the Residual Variance in Regression Analysis. Statist. Neerlandica, 15, 19-23.

TRACY, D.S. and DWYER, P.S. (1969). Multivariate Maxima and Maxima With Matrix Derivatives. J. Amer. Statist. Assoc., 64, 1576-94.

TRACY, D.S. and SINGH, R.P. (1972). A New Matrix Product and its Application in Partitioned Matrix Differentiation. Statist. Neerlandica, 26, 143-57.

TRACY, D.S. and SINGH, R.P. (1975). Some Applications of Matrix Differentation in the General Analysis of Covariance Structures. Sankhya, Series A, 37, 269-280.

VETTER, W.J. (1970). Derivative Operators on Matrices. IEEE Trans. Automatic Control, AC-15, 241-244.

WATSON, G.S. (1964). A Note on Maximum Likelihood. Sankhya, Series A, 26, 303-4.

WONG, Chi Song (To Appear). Matrix Derivatives and Its Applications in Statistics. J. Math. Psychology.

WONG, Chi Song (To Appear). Mathematical Statistics. Tamkang Chair, Tamkang College of Arts and Sciences, Taiwan.

WONG, Chi-Song and WONG, Kai Sang (1979). A First Derivative Test for the Maximum Likelihood Estimates. Bull. Inst. Math. Acad. Sinica, 7, 313-21.

WONG, Chi Song and WONG, Kai Sang* (To Appear). Minima and Maxima in Multivariate Analysis. Canadian J. Statist.

*These references have been suggested by some examiners in the examination committee. They were not available to the author at the time of writing this dissertation.

VITA AUCTORIS

Born	February, 22, 1952	Chung San, China
M.Sc.	May, 1976	University of Windsor
Lecturer (Part Time)	February - July, 1979	National Tsing Hua University, Taiwan

At present, Assistant - Management Sciences, Management Sciences Division, Bell Canada.