1972

SOME GENERALIZATIONS IN MATRIX DIFFERENTIATION WITH APPLICATIONS IN MULTIVARIATE ANALYSIS.

RANA PRATAP. SINGH
University of Windsor

73-6464

SINGH, Rana Pratap, 1943-
SOME GENERALIZATIONS IN MATRIX DIFFERENTIATION
WITH APPLICATIONS IN MULTIVARIATE ANALYSIS.

University of Windsor (Canada), Ph.D., 1972
Statistics

University Microfilms, A XEROX Company, Ann Arbor, Michigan

SOME GENERALIZATIONS IN MATRIX DIFFERENTIATION WITH APPLICATIONS IN MULTIVARIATE ANALYSIS

BY

RANA P. SINGH

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

1972

ABSTRACT

SOME GENERALIZATIONS IN MATRIX DIFFERENTIATION WITH APPLICATIONS IN MULTIVARIATE ANALYSIS

This work deals with certain theoretical generalizations and applications of matrix differential calculus. Various auxiliary and non-auxiliary matrices related to identity matrices and some new matrix products are introduced and their properties are studied. These matrices and a new matrix product known as the partitioned Kronecker product are used to establish interrelationships among column vector representations of matrix functions involving partitioned and non-partitioned matrices. Some particular cases of these new concepts are also pointed out.

In discussing matrix differentiation, it is shown that a variety of results which are true for scalar and vector functions are also true for matrix functions. For example, several algebraic properties of matrix derivative transformations and procedures for identifying various kinds of matrix of partial derivatives and their applications in testing extrema of matrix functions are certain extensions of the corresponding results from ordinary differential calculus. Some interrelationships between four available matrix derivatives are established with the help of a diagram. Differentiation in certain special situations is also presented. These situations are (1) differentiation of matrix functions where scalar values of matrix elements have equality relationships, and (2) differentiation of partitioned

matrix functions. These (matrix differentiation) results are obtained by using the Kronecker matrix product, the Schur product, the partitioned Kronecker product, auxiliary and non-auxiliary matrices and some other algebraic concepts.

Such results are used for estimating matrix parameters in the general multivariate linear and non-linear regression analysis, structural econometric analysis and general covariance structural analysis. Some of the procedures considered here are generalizations of earlier methods in factor analysis. Minimum variance unbiased estimates of the matrix parameter and its linear function are obtained in the general multivariate linear regression model. The asymptotic covariance matrix of unknown parameters in the structural econometric model is also discussed, using partitioned matrix differentiation. Some special matrix derivative formulae are used to obtain the jacobians of the symmetric matrix transformations in a simpler way. A few miscellaneous applications of the new matrix concepts and matrix differentiation formulae are also given. These consist of the derivation of some properties of certain particular matrix products, applications of matrix derivatives to covariance analysis with linear structure, applications to dynamic econometric analysis and derivation of a large sample non-central distribution of a multi-sample regression model with covariance. An up to date listing of various important properties of the Kronecker matrix product, including some new properties, is provided in an appendix. This information is useful for pursuing further investigations concerning the partitioned Kronecker product mentioned above.

ACKNOWLEDGEMENTS

I wish to express my deep feeling of gratitude to my supervisor Dr. D. S. Tracy for suggesting the topic of this dissertation, and for his unfailing help, constant and inspiring guidance, and valuable criticism without which this study would not have been possible.

I am very grateful to Professor P. S. Dwyer of the University of Michigan for his stimulating suggestions regarding this investigation.

It gives me pleasure to thank Dr. A. M. Mathai of McGill University for his many helpful comments.

I am thankful to Professor N. Shklov for his generosity with his time and valuable suggestions on the final draft of this manuscript. I also wish to thank Dr. A. C. Smith and Dr. F. Wm. Lemire for suggesting valuable improvements. In addition, I wish to thank Rev. D. T. Faught for his advice during my graduate work at the University of Windsor.

Finally, I wish to express particular gratitude to Dr. J. P. Clarke of the School of Social Work, University of Windsor, for his encouragement and help.

I acknowledge the support received from the Defence Research Board, National Research Council, National Health Grant, Ontario Government Fellowship and the Department of Mathematics.

TABLE OF CONTENTS

Page

LIST OF TABLES AND FIGURES

Page

TABLE

FIGURE

CHAPTER I

INTRODUCTION

1.1 Review of the Literature on Matrix Differentiation

The principal contributions, in chronological order, to the theory of matrix differentiation, were made by the following authors: Cayley (1845, 1846), Capelli (1887), Turnbull (1927-29), 1930-31a, 1930-31b), Garding (1947), Turnbull (1947), Dwyer & MacPhail (1948), Aitken (1953)*, Wallace (1953), Coy (1955), Wrobleski (1963), Neudecker (1967), Dwyer (1967), Neudecker (1968, 1969b), Tracy & Dwyer (1969) and Gordon & Mathai (1972).

There are two main approaches followed in differentiating matrix functions. These are

(1) using Turnbull's matrix differential operator, which was inspired by the Cayley and Capelli operators,

(2) using the column vector representation of Neudecker and Tracy-Dwyer, which was inspired by Dwyer's (1967) results.

Suitable arrangements of partial derivatives of certain matrix transformations for evaluating their jacobians were first suggested by Deemer & Olkin (1951)#. This technique was

---

* According to Turnbull (1947), the contents of this paper were communicated to him in March 1946.

Taken from lectures of P.L. Hsu delivered in the spring of 1947 at the University of North Carolina.

further extended and simplified by Olkin (1951, 1953), Olkin & Roy (1954), Jack (1964-65) and Olkin & Sampson (1969).

Cayley (1845, 1846) introduced the determinant of the matrix differential operator whose elements were partial differential oeprators of a square matrix of independent variables and studied the properties of this operator. Capelli (1887) considered the minors and linear combinations (polarized forms) of minors of the same order belonging to Cayley's determinantal operator, as described above, and gave a theorem based on his new determinantal operator. Turnbull (1927-29, 1930-31a, 1930-31b) introduced the Ω-operator, which is nothing but Cayley's matrix differential operator, and discussed it

(1) as a useful generalization of differentiation of one scalar variable,

(2) in relation to matrix operands and invariants.

Turnbull established a close analogy between algebraic theory of quantum-numbers, developed by Dirac (1926), and his own differential calculus of finite matrices. He applied the matrix differential operator to three types of elementary functions of characteristic roots of a single matrix variable. He called it the Trace Differential operator. Using this operator, he developed a matrix form of Taylor's Theorem. Simultaneously and independently, Aitken (publication of 1953) and Garding (1947) modified Cayley's operator and extended the general results obtained by him to the case of symmetric matrices. A more general and polarized form of Garding's theorem and a modification of Capelli's theorem, applicable to symmetric and skew-symmetric matrices, were obtained by Turnbull (1947). Wallace (1953) provided an alternative derivation of the modified Capelli Theorem and of Garding's theorem. Further generalization of Cayley and Capelli operators was made by Turnbull (1949). Gordon (1967) used Turnbull's Ω-operator approach to differentiate the multivariate characteristic functions for characterizing the multivariate distributions using regression properties.

Some other methods of differentiating a matrix function were given independently by MacDuffee (1946) and Ferrar (1951).

All the above methods of matrix differentiation are limited either to square matrices or by the necessity for conformability for matrix multiplication of the derivative operator with the matrix function and hence are not of general applicability to multivariate statistical analysis, in which we come across matrix variables of arbitrary dimension.

A matrix differential calculus for scalar functions of a matrix variable and some applications of this calculus to statistics were introduced by Dwyer & MacPhail (1948). Coy (1955) derived the results of Dwyer and MacPhail's paper in a slightly different and simpler manner and extended these results to obtain a differential calculus for scalar and matrix functions of real and complex variables, including functions whose arguments are symmetric, skew-symmetric and diagonal matrices. Wrobleski (1963) provided several extensions of the Dwyer-MacPhail results to more general matrix functions and pointed out their statistical and econometric applications.

Dwyer (1967) extended the concepts of symbolic matrix derivatives, introduced by Dwyer & MacPhail (1948), by developing a general theory of matrix derivatives. He applied his general theory to many multivariate statistical analysis problems, such as maximum likelihood estimation, evaluation of jacobians, generalization of a scalar integral to matrix integral and optimization with side conditions. Also included in Dwyer's (1967) paper are interesting comments about various approaches in the field of matrix differentiation, including his own. For example, the approaches followed in Turnbull (1927-29, 1930-31a, 1930-31b, 1947), Fraser, Duncan & Collar (1936) and Garding (1947) have very few and limited applications. Also Anderson (1958), Wilks (1962) and Rao (1965) presented some results in multivariate statistical theory with the use of scalar differentiation. Thus a rigorous matrix differential calculus may not be very essential to study some aspects of multivariate analysis. On the other hand, Dwyer (1967) justified the fact that a few well-chosen matrix derivative formulae are as essential as an appendix on matrix theory included in many books [Roy (1957), Anderson (1958), and Scheffé (1959) for example], dealing with multivariate statistical theory.

Some matrix differentiation formulae and their applications in obtaining least squares estimates and minimum variance unbiased estimates for ordinary linear regression were presented by Neudecker (1967). Neudecker (1969b) suggested a new method of identifying the matrices of first and second order derivatives, by linking the differentials of the matrix function (involving ordinary and Kronecker matrix products) and the argument matrix. Some of these results were earlier obtained by Dwyer (1967) using a different method. Neudecker (1969b) also applied his method to an econometric problem of maximum likelihood estimation. In the spirit of Dwyer (1967), a further step was taken by Tracy & Dwyer (1969) in obtaining formulae for matrices of first and second order matrix derivatives. They applied their results to problems of extrema of differentiable scalar functions of matrices. This approach simplified the derivation to a great extent. The use of Hessian matrices in making tests for extrema in multivariate analysis was demonstrated by Tracy & Dwyer (1969). The matrix differential calculus presented in Dwyer (1967), Neudecker (1969b), and Tracy & Dwyer (1969) is also useful in evaluating jacobians of some simple matrix transformations occurring in multivariate distribution theory.

Jacobians of certain matrix transformations were obtained by Deemer & Olkin (1951), Olkin (1951, 1953), Olkin & Roy (1954), Roy (1957) and Hua (1958) using elementwise differential techniques, and by Hsu (1953), Hua (1958) and Jack (1964-65) using induction procedures, but in either case without entering into the explicit presentation of matrix derivatives of matrix functions. Both methods are based on a chain of more easily tractable matrix transformations. Khatri (1965) obtained the jacobians of certain complex matrix transformations in order to discuss the statistical analysis of the complex Wishart distribution. Further jacobian results by Khatri (1968) are applicable to the distributions of traces of Wishart matrices.

Dwyer (1958) in dealing with minimum variance unbiased estimation of linear functions of parameter vectors, and Stroud (1968) in discussing Wald statistics, use a technique made available in the paper of Dwyer & MacPhail (1948). Kleinbaum (1970) obtained best asymptotically normal estimates for generalized multivariate linear models. He applied some matrix differentiation formulae given in Dwyer (1967). Mulaik (1971), Gebhardt (1971), Stroud (1971) and Gordon & Mathai (1972) have also presented some applications of matrix differentiation.

1.2 Notation and Abbreviations

Because the shape and properties of matrices are important in finding matrix derivatives and the jacobians of matrix transformations, we adopt the following notation:

1. A^T denotes the transpose of the matrix A, and $A^{-T} = (A^{-1})^T$.
2. Column vectors are denoted by underlined lower case letters, e.g., $\underline{a}^T$:p×1 = $(a_1,\ldots,a_p)$. Matrices are denoted by capital letters.
3. A = (a_{ij}):p×q means that the matrix A has p rows and q columns with a_{ij} as the element in the i-th row and j-th column.
4. I:p×p or I_p denotes a p×p identity matrix.
5. O:p×q and $\underline{O}$:p×1 denote a p×q matrix of zeros and a p×1 vector of zeros respectively.
6. tr A denotes the trace of a square matrix A = (a_{ij}) (that is, the sum of the leading diagonal elements $= \sum_i a_{ii}$).
7. |A| denotes the determinant of A.
8. A^{-1} denotes the unique inverse of a non-singular square matrix A.
9. A^- is a generalized inverse (g-inverse) of A, defined implicitly as any A^- satisfying $AA^-A = A$ for any A.
10. A⊗B:pr×qs denotes the Kronecker product of A:p×q, B:r×s defined as $A \otimes B = (a_{ij}B)$.

11. $A{\times}{\times}B{:}p{\times}q$ denotes the Schur (Hadamard) product of $A{:}p{\times}q$, $B{:}p{\times}q$ defined as $A{\times}{\times}B = (a_{ij}b_{ij})$.

12. $X_r{:}pq{\times}1$ ($X_c{:}pq{\times}1$) denotes the column vector representation of the matrix elements of $X{:}p{\times}q$ by displaying underneath each other the rows (columns) of X by taking in order the first row, then the second row, etc., (first column, then the second column etc.). See Section 2.6.

13. $I_{(k)}$, for a suitable integer k, is a permuted identity matrix. See Section 2.4.

14. $dA = (da_{ij}){:}p{\times}q$ denotes the differential of $A{:}p{\times}q$.

15. We use the word "conformability of matrices" to mean that various partitioned and non-partitioned matrix operations (for example sum, difference, multiplication, Schur product and a few other matrix products) are well defined. For instance $A{\times}{\times}B{:}p{\times}q$ exists only if $A{:}p{\times}q$ and $B{:}p{\times}q$.

16. $\circ$ denotes the composition of two matrix functions.

17. $X^{-k} = (X^{-1})^k$, where k is an integer.

18. $\langle X \rangle$ or $\langle x_{ij} \rangle$ denotes the (i,j)-th matrix element of X.

1.3 Aim of the Present Study

This dissertation is concerned with some generalizations of the theory and techniques of the method of matrix differentiation proposed by Neudecker (1969b), and Tracy & Dwyer (1969). These generalizations are discussed in Chapters Two and Three. Most results are applicable to estimation problems in multivariate models. Some of these applications are given in Chapter Four.

In Chapter Two a more general definition of a matrix function is proposed. Based on this, various types of matrix functions, some of which occur in statistical applications, are discussed. Some auxiliary and non-auxiliary matrices related to non-partitioned and partitioned identity matrices are introduced. Three new matrix products, one of which we call the partitioned Kronecker product, are defined. These products have various interesting properties. Some of the useful results concerning the Kronecker product of matrices and the partitioned Kronecker product are obtained by using the above mentioned auxiliary matrices. These results are applicable in establishing a very general method of column vector representation and in deducing the various interrelationships of the resulting representations of matrix functions involving partitioned and non-partitioned matrices.

In Chapter Three some general results concerning matrix differentiation are presented. In this chapter a general definition of differentiability of a matrix function is proposed. In addition to having many other properties which the matrix of partial derivatives possesses, it is looked upon as a matrix of certain linear transformation. Some results concerning basis representations of the partial matrix derivatives are established. For any complicated matrix function, a general procedure for identifying the matrix of partial derivatives is also discussed. It is suggested that the identification of only one form of partial matrix derivative out of the four available forms in this work is enough for the purpose of statistical applications because there exist natural relationships among these forms. A theorem for identification of mixed partial matrix derivatives for differentiable scalar functions of two matrix variables is established. Certain corollaries are also given. This theorem extends a result of Neudecker (1969b). Some results for testing the unconstrained and constrained extrema of matrix functions are obtained. These are generalizations of the corresponding results for vector functions given in Goldberger (1964) and Tracy & Dwyer (1969). Further generalizations lead to the introduction of matrix differentiation (1) after partitioning matrices into blocks and (2) with equality relationships among scalar values of matrix elements. Non-partitioned matrix differentiation uses auxiliary and non-auxiliary matrices, the ordinary Kronecker product and its various properties. The study of partitioned matrix differentiation is made simpler by applying various extended concepts of non-partitioned situations, the partitioned Kronecker product and its properties.

Some applications of the matrix differential calculus, presented in Chapter Three, are given in Chapter Four. Linear and non-linear parameter estimation in multivariate regression models is discussed. These discussions provide some extensions of the results in Tan (1968-69) and Allen (1967). Minimum variance unbiased estimation, discussed by Dwyer (1958), is extended to a general regression model. Estimation of parameters and their asymptotic covariances in systems of simultaneous equations are discussed, using the results from partitioned matrix differentiation. Estimating equations for covariance structural parameters are obtained for three methods of covariance structural analysis. For the first two methods, Hessian matrices are also obtained. For a general covariance structural model, an expression for estimating unobserved variables is given. This expression uses basic results from partitioned matrices. A modified matrix derivative is used to obtain the jacobians of certain symmetric matrix transformations in a simpler manner. These jacobians are useful in evaluating matrix integrals which are applied in deriving multivariate distributions. Some miscellaneous applications include the derivation of properties of matrix products defined by Khatri & Rao (1968) and Khatri (1971), estimation in covariance analysis with linear structure, in dynamic econometric analysis and in the derivation of the asymptotic non-null distribution of the multi-sample regression model with covariance. The use of various matrix concepts from Chapter Two is made in the simplification of certain calculations of these results.

Some suggestions for further research are given in Chapter Five.

A few important results on matrices, some of which are new, are given in the Appendix.

The whole development is over the field of real numbers, though most of the results would hold over the field of complex numbers.

CHAPTER II

FUNCTIONS OF MATRICES

2.1 Introduction

The notation $\underline{y} = \underline{f}(\underline{x})$ is well known in differential calculus as representing a vector valued function of a vector variable. For a scalar variable x, a few results on differentiation and integration of matrix functions of the type $Y = Y(X(x))$, where X is a square matrix, are available in the work of Michal (1947). Neudecker (1969b) and Tracy & Dwyer (1969) have presented a matrix differential calculus for matrix functions of one matrix variable. However, they have not presented a matrix differential calculus for more general matrix functions of several matrix variables. In this chapter we provide various matrix functions, some of which are presently unknown, but are needed for completeness in the general matrix differential calculus. Some new matrices and some new matrix products are introduced and their properties are studied. Some other results concerning matrix functions and the Kronecker matrix product are also obtained.

2.2 Definition of a Matrix Function

Let $\mathcal{M}_{p,q}(R)$ and $\mathcal{N}_{m,n}(R)$ be the spaces of p×q matrices and m×n matrices respectively over the field R of real numbers, then the mapping

$$F : \mathcal{M}_{p,q}(R) \longrightarrow \mathcal{N}_{m,n}(R)$$

defined by

$$F_Y(X) = Y(X) = \begin{bmatrix} y_{11}(X) & \cdots & y_{1n}(X) \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ y_{m1}(X) & \cdots & y_{mn}(X) \end{bmatrix} \tag{2.2.1}$$

is called a <u>matrix function of a matrix variable</u> $X = (x_{ij})$, where $y_{\alpha\beta}(X)$ and x_{ij} are in R, for each $\alpha = 1,\ldots,m$, $\beta = 1,\ldots,n$; and $i = 1,\ldots,p$, $j = 1,\ldots,q$.

<u>Example 2.1</u> Consider the matrix function

$$Y = AX^TB \tag{2.2.2}$$

where A:m×q, B:p×n are matrices of constant scalars, and $X \in \mathcal{M}_{p,q}(R)$, $Y \in \mathcal{N}_{m,n}(R)$.

<u>Example 2.2</u> Consider, for A and X as in Example 2.1,

$$Y = A \otimes X^TX \tag{2.2.3}$$

where ⊗ is the Kronecker product, and $Y \in \mathcal{N}_{mq,q^2}(R)$.

<u>Example 2.3</u> Let ×× be the Schur product. Then for a square matrix A:p×p and a non-singular matrix X:p×p, we have

$$Y = A \times\times X^{-1} \tag{2.2.4}$$

where $Y \in \mathcal{N}_{p,p}(R)$.

The matrix function of a vector variable and the matrix function of a scalar variable are obtained as particular cases of the above definition by considering X as a vector and as a scalar respectively. Similarly the vector function and scalar function of a matrix variable (also of a vector variable, a scalar variable) are particular cases of a matrix function of a matrix variable, defined above, obtained by considering Y as a vector and as a scalar respectively.

Example 2.4 A mapping I defined by

$$I(X) = X{:}p\times q \tag{2.2.5}$$

is called an identity matrix function of a matrix variable X from $\mathscr{M}_{p,q}(R)$ into itself.

Situations may arise where, in the expression (2.2.1), the matrix variable X is a function of a matrix variable $Z \in \mathscr{L}_{r,s}$, i.e.,

$$X = G(Z). \tag{2.2.6}$$

Then

$$Y = F(X) = F(G(Z)) = (F\circ G)(Z) \tag{2.2.7}$$

where Y is a matrix function of a matrix variable Z from $\mathscr{L}_{r,s}(R)$ into $\mathscr{N}_{m,n}(R)$.

The matrix function of more than one matrix variable is defined as a mapping from the cartesian product of spaces as domain into a space as range. In particular, we define below a matrix function of two matrix variables.

Let $X \in \mathscr{L}_{p,q}(R)$, $Z \in \mathscr{M}_{r,s}(R)$. Then a mapping

$$G : \mathscr{L}_{p,q}(R) \times \mathscr{M}_{r,s}(R) \longrightarrow \mathscr{N}_{m,n}(R)$$

defined by

$$G_Y(X,Z) = Y(X,Z) \tag{2.2.8}$$

is a matrix function of two matrix variables X, Z; where $Y \in \mathscr{N}_{m,n}(R)$.

Example 2.5 Consider

$$Y = AX^TBZC \tag{2.2.9}$$

where A:m×q, B:p×r, and C:s×n are non-variable matrices;

$Y \varepsilon \mathcal{N}_{m,n}(R)$ is a dependent matrix variable, and

$X \varepsilon \mathcal{L}_{p,q}(R)$, $Z \varepsilon \mathcal{M}_{r,s}(R)$ are independent matrix variables.

Various other examples may similarly be given.

2.3 Types of Function

We shall represent a more general matrix function by

$$Z = F(X, Y, \ldots, W) \qquad (2.3.1)$$

where X, Y,..., W are independent variable matrices, Z is a dependent variable matrix and F represents a matrix-valued function. Various particular cases of Z may be considered. For example, consider the following functions where the matrices involved are conformable:

$$Z = AX^rB(Y^T)^sC(X^T)^tDY^u, \quad r, s, t, u \text{ positive integers} \qquad (2.3.2)$$

$$Z = (a_o I + a_1X + \cdots + a_pX^p)(b_o I + b_1X + \cdots + b_qX^q)^{-1}, \quad p, q \text{ positive integers} \qquad (2.3.3)$$

$$Z = e^{F(X, Y, \ldots, W)} \quad \text{where } F \text{ is a square matrix} \qquad (2.3.4)$$

$$Z = \text{Log } F(X) \quad \text{where } F \text{ is a square matrix} \qquad (2.3.5)$$

$$Z = X^{\frac{r}{s}}, \quad \text{where } r, s \text{ are positive integers} \qquad (2.3.6)$$

$$Z = X \otimes Y \quad \text{where } \otimes \text{ is the Kronecker matrix product} \qquad (2.3.7)$$

$$Z = X{\times\times}Y{\times\times}W, \quad {\times\times} \text{ is the Schur product} \qquad (2.3.8)$$

$$\underline{z} = A\underline{x} \qquad (2.3.9)$$

$$z = |AX^rB(Y^T)^s|^t, \quad r, s, t \text{ positive integers} \qquad (2.3.10)$$

$$z = |X-\lambda I| \quad \text{where } \lambda \text{ is a constant scalar} \qquad (2.3.11)$$

$$z = \text{tr } F(X, Y, \ldots, W), \text{ where } F \text{ is a square matrix.} \qquad (2.3.12)$$

Some other functions may similarly be constructed. Thus, in general, there are various types of functions of matrices, each of which can be divided into several categories.

A few types of functions of matrices are the following:

(1) matrix functions of several matrix arguments together with scalar coefficients;

(2) matrix functions of several matrix arguments together with matrix coefficients;

(3) scalar functions of several matrix arguments together with scalar coefficients.

Each of the above types of functions of matrices could take one of the following forms:

(1) that involving ordinary matrix products;

(2) that involving Schur products;

(3) that involving Kronecker matrix products;

etc. New matrix products $\odot$ and ⓔ, defined by Khatri & Rao (1968) and Khatri (1971) respectively, and $\textcircled{\pi}$ defined in Section 2.7 of this dissertation, yield three additional categories of matrix functions.

The following are certain commonly occurring matrix functions, some of which are applicable in multivariate analysis:

$$Y = X^{-1} \tag{2.3.13}$$

$Y = X^{-}$ is a relation involving a g-inverse X^{-} of X (2.3.14)

$Y = BXB^{T}$, where $B{:}p\times p$ is a non-singular upper triangular matrix (2.3.15)

$$\text{Sin } X = \sum_{\alpha=0}^{\infty} (-1)^{\alpha} \frac{X^{2\alpha+1}}{(2\alpha+1)!}, \text{ X is a square matrix} \quad (2.3.16)$$

$$\text{Cos } X = \sum_{\alpha=0}^{\infty} (-1)^{\alpha} \frac{X^{2\alpha}}{(2\alpha)!}, \text{ X is a square matrix} \quad (2.3.17)$$

$$e^{X} = \sum_{\alpha=0}^{\infty} \frac{X^{\alpha}}{\alpha!}, \text{ X is a square matrix} \quad (2.3.18)$$

$$\text{Log } (I-X) = -\sum_{\alpha=0}^{\infty} \frac{X^{\alpha}}{\alpha}, \text{ X is a square matrix} \quad (2.3.19)$$

$$\begin{aligned} \log |I-X| &= \text{tr Log } (I-X), \text{ X is a square matrix} \\ &= -\text{tr}[X + \frac{X^2}{2} + \frac{X^3}{3} + \cdots] \end{aligned} \quad (2.3.20)$$

$h(W) = |W|$, where W is the residual moment matrix (used in econometrics) (2.3.21)

$h(W) = \text{tr } S^{-1}W$, S is the residual moment matrix of unconstrained regression (2.3.22)

$g(\beta) = \text{tr } \Sigma^{-1}(X-AF(\beta))^T V^{-1}(X-AF(\beta))$, where $F(\beta)$ is some matrix function of β (2.3.23)

$f(X) = (2\pi)^{-\frac{pn}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\frac{n}{2} \text{tr } \Sigma^{-1}T}$, where $X{:}n \times p$, $E(X) = A\beta P$, and rows of X are independently and normally distributed with the same covariance matrix Σ and

$$T = n^{-1}(X-A\beta P)^T (X-A\beta P). \quad (2.3.24)$$

Matrix derivative results for some of the above functions are presented in Chapter III.

2.4 Some Special Types of Matrices

In this section we introduce the following five types of matrices representing operations which are related to an identity matrix. The first two are called auxiliary matrices and the last three are called non-auxiliary matrices. The auxiliary matrices preserve the dimension and the number of 1's and 0's of the basic identity matrix, whereas the non-auxiliary ones do not.

Definition 2.4.1 (Tracy & Dwyer (1969)). Suppose we rearrange the rows of an identity matrix $I: mn\times mn$ by taking every m-th row (n-th row) starting with the first, then every m-th row (n-th row) starting with the second etc. This matrix we denote by $I_{(m)}$ $(I_{(n)})$.

Definition 2.4.2 Suppose we rearrange the columns of an identity matrix $I: mn\times mn$ by taking every m-th column (n-th column) starting with the first, then every m-th column (n-th column) starting with the second etc. This matrix we denote by $I^{(m)}$ $(I^{(n)})$.

Example 2.4.1 Let $m = 3$, $n = 2$. Then

$$I_{(3)} = \begin{bmatrix} 1&0&0&0&0&0\\ 0&0&0&1&0&0\\ 0&1&0&0&0&0\\ 0&0&0&0&1&0\\ 0&0&1&0&0&0\\ 0&0&0&0&0&1 \end{bmatrix}, \quad I^{(3)} = \begin{bmatrix} 1&0&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&0&1&0\\ 0&1&0&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&0&1 \end{bmatrix}$$

$I_{(2)} = I^{(3)}$ and $I^{(2)} = I_{(3)}$.

In general we may verify the following results:

Theorem 2.4.1 For a given identity matrix $I: mn\times mn$, we have

$$I_{(m)}I_{(n)} = I_{(n)}I_{(m)} = I \tag{2.4.1}$$

$$I_{(m)} = I_{(n)}^{T} \tag{2.4.2}$$

$$I_{(m)} = I^{(m)^T} = I^{(n)} \ , \quad I_{(n)} = I^{(m)} = I^{(n)^T} \qquad (2.4.3)$$

$$I_{(m)} = I_{(n)} \ \text{ if } m = n \qquad (2.4.4)$$

$$I_{(m)} + I_{(n)} \text{ is a symmetric matrix} \qquad (2.4.5)$$

$$|I_{(m)}| = |I_{(n)}| = \pm 1 \qquad (2.4.6)$$

In the later part of this work, we shall not require the second type of auxiliary matrices (Definition 2.4.2) because these turn out to be simple transformations of the first type of auxiliary matrices. This can be noticed from Theorem 2.4.1 (2.4.2) and (2.4.3) above. Some of the results above go back to Tracy & Dwyer (1969).

In the following discussion k and ℓ are the number of repeated scalar elements in the matrices Y and X respectively, where Y = F(X). Specific positions of these elements in the matrix lead to deleting certain rows from (or appending certain rows to) to an identity matrix. Some applications of these concepts are provided in Sections 3.11 and 4.7.

Definition 2.4.3 Suppose we delete a particular set of k rows, other than the first row, from an identity matrix I:pq×pq. We denote this matrix by M:(pq−k)×pq. (These k rows correspond to the k repeated elements.)

Definition 2.4.4 We denote by N:mn×(mn−ℓ) a matrix obtained by appending to the identity matrix of order mn−ℓ a particular set of ℓ of its rows other than the first. These rows may either follow or be inserted between the rows of the identity matrix. (These ℓ rows correspond to the ℓ repeated elements.)

Example 2.4.2 Take $p = m = 3$, $q = n = 2$, $k = \ell = 1$. Suppose further that we delete fifth row from I_6 and append third row of I_5 between its fourth and fifth rows. Then

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad N = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Definition 2.4.5 We denote by $I^*:mn\times mn$ a matrix obtained by replacing a particular set of the 1's in the identity matrix $I:mn\times mn$ by zeros. (The choice of the elements so replaced is governed by the problem under consideration, as in Section 3.11.)

Under a particular approach, the matrices M, N and I^* are uniquely obtained.

Example 2.4.3 Let $m = 2$, $n = 3$. Then replacing the first and fifth 1's on the diagonal of I_6 by 0's, we get

$$I^* = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Theorem 2.4.2 For a given $I:mn\times mn$, the following basic properties are immediately evident.

$$MM^T = I \tag{2.4.7}$$

$$M^TM = I^* \tag{2.4.8}$$

$$NN^T = I + \sum_{i}\sum_{\substack{j \\ i\neq j}}(E_{i,j} + E_{j,i}), \quad i,j \in \{2,\dots,mn\}$$

where $E_{r,s}$ is a square matrix with the (r,s)-th entry as 1 and zeros elsewhere. (2.4.9)

$$N^T N = I + \sum_i E_{i,i} \tag{2.4.10}$$

The determination of the matrix derivatives and the jacobians of certain matrix transformations is made simpler by investigating some procedures which are based on the matrices $I_{(m)}$, $I_{(n)}$, M, N, and I*.

2.5 Some Equalities Concerning the Kronecker Product

The following theorem is based on the auxiliary matrices $I_{(m)}$, $I_{(n)}$ and enables us to simplify the resulting matrices of partial derivatives.

Theorem 2.5.1 If A and B are p×q and r×s matrices respectively, then

(i) $I_{(r)}(A \otimes B) = (B \otimes A) I_{(s)}$ (2.5.1)

(ii) $I_{(q)}(B \otimes A^T) = (A^T \otimes B) I_{(p)}$ (2.5.2)

(iii) $I_{(s)}(A \otimes B^T) I_{(q)} = B^T \otimes A$. (2.5.3)

Proof: (i) We note that the j-th column of $(B \otimes A) I_{(s)}$ is $B_{.\delta} \otimes A_{.\beta}$, where $j = s(\beta-1)+\delta$, $\beta = 1,\ldots,q$, $\delta = 1,\ldots,s$. Also for this j, the j-th column of $A \otimes B$ is $A_{.\beta} \otimes B_{.\delta}$. Since $I_{(r)}(A_{.\beta} \otimes B_{.\delta}) = B_{.\delta} \otimes A_{.\beta}$, the result follows.

(ii) Follows from (i).

(iii) From (i) we get

$$I_{(s)}(A \otimes B^T) I_{(q)} = (B^T \otimes A) I_{(r)} I_{(q)}$$

$$= B^T \otimes A \quad \text{using (2.4.1).}$$

A few other equalities concerning Kronecker product of matrices are available in the Appendix.

2.6 Column Vector Representations of Matrix Functions

The concept of a column vector function of a matrix function in terms of its column vector variables is basic to this dissertation. For any matrix, two types of column vector representations, together with their interrelations, are provided. These are then carried on to general matrix functions of matrix variables, which are useful in identifying the required partial matrix derivatives. Theorem 2.4.1 is used in proving some of the theorems of the present section.

2.6.1 Column Vector Representations of a Matrix

Definition 2.6.1 For a matrix $Y{:}m\times n$, we define its column vector representations Y_r and Y_c as

$$Y_r = \begin{bmatrix} Y_{1.}^T \\ Y_{2.}^T \\ \cdot \\ \cdot \\ \cdot \\ Y_{m.}^T \end{bmatrix}, \qquad Y_c = \begin{bmatrix} Y_{.1} \\ Y_{.2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{.n} \end{bmatrix},$$

where $Y_{i.}^T$ and $Y_{.j}$ are the transpose of the i-th row vector, and the j-th column vector, respectively.

The idea of Y_c is due to Koopmans (1950) (who denoted it by Y_{vec}) and that of Y_r is due to Tracy & Dwyer (1969).

2.6.2 Interrelationships of the Column Vector Representations

The basic interrelations of Y_r and Y_c are given by the following result which enables us to use the auxiliary matrices. From now on we assume that the matrices involved are conformable for matrix operations.

Theorem 2.6.2.1 If Y is an m×n matrix, then

(i) $I_{(m)}Y_c = Y_r$ (2.6.2.1)

(ii) $Y_c = I_{(n)}Y_r$. (2.6.2.2)

These results follow trivially from the definitions. For example, (i) follows from the definition of $I_{(m)}$ given in Section 2.4.1, since the i, i+m,..., i+(n-1)m-th elements of Y_c form the i-th subvector of $I_{(m)}Y_c$ and also the i-th row of Y, for i = 1, 2,..., m. The fact that the i-th row of Y is the i-th subvector of Y_r proves (i).

Reformulation of the proof for (i), with the roles of rows and columns interchanged, establishes (ii).

The basic property (2.4.1) follows from Theorem 2.6.2.1. In fact, by using (2.6.2.1) and (2.6.2.2), we get

$$\left.\begin{array}{l} IY_r = Y_r = I_{(m)}Y_c = I_{(m)}I_{(n)}Y_r \\ \\ IY_c = Y_c = I_{(n)}Y_r = I_{(n)}I_{(m)}Y_c \end{array}\right\} \Longrightarrow I_{(m)}I_{(n)} = I_{(n)}I_{(m)} = I \ .$$

2.6.3 A General Procedure for Column Vector Representations of Matrix Functions

Consider any matrix function Y = F(X). Let $\langle y_{\alpha\beta}\rangle$ denote the matrix value of the (α,β)-th element of Y:m×n, (see Dwyer (1967)). Based on this concept, the vectors Y_r, Y_c whose respective subvectors are

$$Y_{\alpha.} = [y_{\alpha 1} \;\; y_{\alpha 2} \;, \dots, \; y_{\alpha n}] \quad \text{and} \quad Y_{.\beta} = \begin{bmatrix} y_{1\beta} \\ y_{2\beta} \\ \cdot \\ \cdot\cdot \\ \cdot \\ y_{m\beta} \end{bmatrix}$$

are uniquely defined. The vectors X_r and X_c are similarly defined. Now a general method of representing Y_r or Y_c as functions of X_r or X_c or both is basic to this dissertation both from the point of view of the theory and application. For example, we require $Y_r = G(X_r)$ to develop the general matrix differential calculus in Chapter III and to derive some results concerning the minimum variance unbiased estimation, given in Section 4.4, in matrix form. Neudecker (1969b) presented some theorems connecting Y_c and tr Y with the Kronecker product of matrices, Y being ordinary product of conformable matrices. He used these theorems in differentiating, up to second degree, non-linear matrix functions involving ordinary matrix products and Kronecker matrix products, and applied the resulting matrix differentiation to the estimation of the structural parameters in a simultaneous linear structural equation model. In this subsection we develop some theorems which are applied below to identify the matrix of partial derivatives for more general matrix functions in a simpler manner.

We develop the general procedure of column vector representations of matrix functions in the following way:

<u>Theorem 2.6.3.1</u> If $Y = AX$, where $Y{:}m\times n$, $A{;}m\times p$ and $X{:}p\times n$, then

(i) $$Y_r = \sum_j (A_{.j} \otimes X_{j.}{}^T) \text{ where } X_{j.}{}^T = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \cdot \\ \cdot \\ \cdot \\ x_{jn} \end{bmatrix} \qquad (2.6.3.1)$$

(ii) $$Y_r = (I \otimes X^T) A_r \ . \qquad (2.6.3.2)$$

<u>Proof</u>: (i) We have

$$Y_{i.}{}^T = \sum_{j=1}^{p} a_{ij} X_{j.}{}^T \quad ,$$

which gives

$$Y_r = \begin{bmatrix} \sum_{j=1}^{p} a_{1j} X_{j.}{}^T \\ \sum_{j=1}^{p} a_{2j} X_{j.}{}^T \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{p} a_{mj} X_{j.}{}^T \end{bmatrix} = \sum_{j=1}^{p} (A_{.j} \otimes X_{j.}{}^T) \ .$$

(ii) In this case

$$Y_{i.}{}^T = \sum_{j=1}^{p} a_{ij} X_{j.}{}^T$$

$$= (e_{i.} \otimes X^T) A_r \quad , \ e_{i.} = (0,\ldots,\ 0\ \underset{\substack{\downarrow \\ \text{i-th entry}}}{1}\ 0,\ldots,0).$$

Hence

$$Y_r = \begin{bmatrix} e_{1.} \otimes X^T \\ \cdot \\ \cdot \\ \cdot \\ e_{m.} \otimes X^T \end{bmatrix} A_r = (I \otimes X^T) A_r \ .$$

As a consequence of the above theorem, we have

$$Y_r = (A \otimes X^T) I_r \qquad (2.6.3.3)$$

$$= (A \otimes I) X_r \; . \qquad (2.6.3.4)$$

<u>Theorem 2.6.3.2</u> If Y be as in Theorem 2.6.3.1, then

$$\text{(i)} \quad Y_c = \sum_j (X_{j.}^{\;T} \otimes A_{.j}) \qquad (2.6.3.5)$$

$$\text{(ii)} \quad Y_c = (I \otimes A) X_c \; . \qquad (2.6.3.6)$$

<u>Proof</u>: Using Theorems 2.6.2.1 and 2.6.3.1, we get

$$\text{(i)} \quad Y_c = I_{(n)} Y_r$$

$$= I_{(n)} \sum_j (A_{.j} \otimes X_{j.}^{\;T})$$

$$= \sum_j I_{(n)} (A_{.j} \otimes X_{j.}^{\;T}) I_{(1)}$$

$$= \sum_j (X_{j.}^{\;T} \otimes A_{.j}) \text{ , using Theorem 2.5.1(iii).}$$

$$\text{(ii)} \quad Y_c = I_{(n)} (A \otimes I) X_r$$

$$= (I \otimes A) I_{(n)} X_r$$

$$= (I \otimes A) X_c \; .$$

Following results follow readily from Theorem 2.6.3.2:

$$Y_c = (X^T \otimes I) A_c \qquad (2.6.3.7)$$

$$= (X^T \otimes A) I_c \; . \qquad (2.6.3.8)$$

An alternate proof of Theorem 2.6.3.2 is given by Neudecker (1969b).

An immediate consequence of Theorems 2.6.3.1 and 2.6.3.2 is that whenever we have products of several matrices occurring in $Y = F(X)$, we can form Y_r and Y_c without going into tedious calculations. Additional simplifications will be obtained by using the following result:

If X is any m×n matrix, we have

$$(X^T)_c = X_r = I_{(m)}X_c \tag{2.6.3.9}$$

$$(X^T)_r = X_c = I_{(n)}X_r \ . \tag{2.6.3.10}$$

As an example, for $Y = AX^TB$, we have

$$Y_r = (A \otimes B^T)I_{(n)}X_r \tag{2.6.3.11}$$

and

$$Y_c = (B^T \otimes A)I_{(m)}X_c \ . \tag{2.6.3.12}$$

Obviously, for $Y = AXB$, we obtain

$$Y_r = (A \otimes B^T)X_r \tag{2.6.3.13}$$

and

$$Y_c = (B^T \otimes A)X_c \ . \tag{2.6.3.14}$$

Equation (2.6.3.14) is also available in Neudecker (1969b).

It may be noticed that expressions connecting Y_r (and Y_c) with A_r, B_r, and I_r (and A_c, B_c, and I_c) can be easily derived.

Now we consider matrix functions involving integral powers of a matrix variable.

<u>Theorem 2.6.3.3</u> For $Y = X^s$, where X is a square matrix and s is a positive integer, we have

$$\text{(i)} \quad Y_r = \frac{1}{s} \sum_{j=0}^{s-1} [X^j \otimes (X^T)^{s-j-1}]X_r \tag{2.6.3.15}$$

$$\text{(ii)} \quad Y_c = \frac{1}{s} \sum_{j=0}^{s-1} [(X^T)^{s-j-1} \otimes X^j]X_c \ . \tag{2.6.3.16}$$

Proof: (i) Follows by induction. The formula is obvious for $s = 1$. Now suppose that it is true for $Z = X^s$, then for $Y = X^{s+1}$ we have

$$Y_r = (X^s \otimes I)X_r \tag{2.6.3.17}$$

and

$$Y_r = (I \otimes X^T)Z_r$$

$$= \frac{1}{s}(I \otimes X^T) \sum_{j=0}^{s-1} [X^j \otimes (X^T)^{s-j-1}]X_r$$

$$= \frac{1}{s} \sum_{j=0}^{s-1} [X^j \otimes (X^T)^{s-j}]X_r \ . \tag{2.6.3.18}$$

Combination of (2.6.3.17) and (2.6.3.18) yields :

$$Y_r = \frac{1}{s+1} \sum_{j=0}^{s} [X^j \otimes (X^T)^{(s+1)-j-1}]X_r \ .$$

This concludes the proof of (i).

The proof of (ii) is analogous to that of (i) and is therefore omitted.

Theorem 2.6.3.4 For $Y = (X^T)^s$, where $X{:}n{\times}n$ and s is a positive integer, we have

$$\text{(i)} \quad Y_r = \frac{1}{s} \sum_{j=0}^{s-1} [(X^T)^j \otimes X^{s-j-1}]I_{(n)}X_r \tag{2.6.3.19}$$

$$\text{(ii)} \quad Y_c = \frac{1}{s} \sum_{j=0}^{s-1} [X^j \otimes (X^T)^{s-j-1}]I_{(n)}X_c \ . \tag{2.6.3.20}$$

Proof: The proof again follows from (2.6.3.9) and (2.6.3.10) and method of induction, as in Theorem 2.6.3.3.

The procedure adopted in proving Theorems 2.6.3.3 and 2.6.3.4 can be used in proving theorems involving positive or negative integral or rational powers of a matrix variable whenever they are defined. In particular, various matrix functions involving the ordinary matrix product, some of which are listed in Section 2.3, may be represented as vector functions.

It is possible to generalize these theorems to partitioned matrix functions.

2.7 The Partitioned Kronecker Product and its Properties

In Section 2.6 we have seen that the Kronecker product $\otimes$ may be used for the column vector representations Y_r, Y_c of ordinary matrix functions $Y = F(X)$, in terms of X_r and/or X_c. Sometimes we come across matrix functions where some or all of the matrices involved are partitioned into submatrices. To obtain the column vector representations of partitioned matrix functions, we need to extend the idea of Kronecker product $A \otimes B$ of two matrices A and B, as stated in Definition A.1.1.

Definition 2.7.1 Let

$$A = \begin{matrix} & k_1 \; \cdots \; k_q \\ \begin{matrix} m_1 \\ \vdots \\ m_p \end{matrix} & \begin{bmatrix} \\ \quad A^{ij} \quad \\ \\ \end{bmatrix} \end{matrix}, \quad B = \begin{matrix} & t_1 \; \cdots \; t_h \\ \begin{matrix} s_1 \\ \vdots \\ s_g \end{matrix} & \begin{bmatrix} \\ \quad B^{uv} \quad \\ \\ \end{bmatrix} \end{matrix}$$

be two matrices which are partitioned arbitrarily. Then the partitioned Kronecker product $A \circledS B$ (which may be read as A pi B) is defined by

$$A \circledast B = \begin{bmatrix} (A^{11} \# B) & \cdots & (A^{1q} \# B) \\ \vdots & & \\ (A^{p1} \# B) & \cdots & (A^{pq} \# B) \end{bmatrix}$$

$$= [A^{ij} \# B] \; ,$$

where

$$A^{ij} \# B = \left[\begin{array}{c|c|c} A^{ij} \otimes B^{11} & \cdots & A^{ij} \otimes B^{1h} \\ \hline \vdots & & \\ \hline A^{ij} \otimes B^{g1} & \cdots & A^{ij} \otimes B^{gh} \end{array}\right]$$

and $A^{ij} \otimes B^{uv}$ is the well-known Kronecker product $(a^{ij}_{\alpha\beta} \otimes B^{uv})$.

$A \circledast B$ may be looked upon as a partitioned matrix of order $ms \times kt$, where $m = \sum_{i=1}^{p} m_i$, $k = \sum_{j=1}^{q} k_j$, $s = \sum_{u=1}^{g} s_u$, and $t = \sum_{v=1}^{h} t_v$.

Example 2.7.1 Let

$$A = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}, \quad B = \begin{bmatrix} B^{11} & B^{12} & B^{13} \\ B^{21} & B^{22} & B^{23} \end{bmatrix}$$

then

$$A \circledast B = \begin{bmatrix} A^{11} \otimes B^{11} & A^{11} \otimes B^{12} & A^{11} \otimes B^{13} & A^{12} \otimes B^{11} & A^{12} \otimes B^{12} & A^{12} \otimes B^{13} \\ A^{11} \otimes B^{21} & A^{11} \otimes B^{22} & A^{11} \otimes B^{23} & A^{12} \otimes B^{21} & A^{12} \otimes B^{22} & A^{12} \otimes B^{23} \\ A^{21} \otimes B^{11} & A^{21} \otimes B^{12} & A^{21} \otimes B^{13} & A^{22} \otimes B^{11} & A^{22} \otimes B^{12} & A^{22} \otimes B^{13} \\ A^{21} \otimes B^{21} & A^{21} \otimes B^{22} & A^{21} \otimes B^{23} & A^{22} \otimes B^{21} & A^{22} \otimes B^{22} & A^{22} \otimes B^{23} \end{bmatrix}.$$

Many properties of the partitioned Kronecker product are similar to those of the Kronecker product (Graybill (1969, pp. 197-209)). We assume that the matrices involved are conformable with respect to partitioned addition and partitioned multiplication. The following results involving the partitioned Kronecker product are straightforward:

$A \circledast B,\ B \circledast A$ exist for any A, B (2.7.1)

$A \circledast B \neq B \circledast A$ in general (2.7.2)

$(A \circledast B) \circledast C = A \circledast (B \circledast C)$ (2.7.3)

$(\alpha A) \circledast B = A \circledast (\alpha B) = \alpha(A \circledast B)$, where α is a scalar (2.7.4)

$(A \circledast B)^T = A^T \circledast B^T$ (2.7.5)

$(A \circledast B)(C \circledast D) = AC \circledast BD$ if AC, BD exist (2.7.6)

$(A+B) \circledast (C+D) = A \circledast C + A \circledast D + B \circledast C + B \circledast D$ (2.7.7)

$(A \circledast B)^{-1} = A^{-1} \circledast B^{-1}$ if A^{-1}, B^{-1} exist (2.7.8)

$I_m \circledast I_t = I_{mt}$ for identity partitioned matrices (2.7.9)

$(A \circledast I_n)(I_m \circledast B) = (I_m \circledast B)(A \circledast I_n) = A \circledast B$, where

A:m×n, B:n×n, I_m and I_n are partitioned matrices (2.7.10)

tr $(A \circledast B)$ = (tr A)(tr B) if traces exist. (2.7.11)

Analogues of (A.1.20)-(A.1.26)* are also true for $\circledast$.

Some other properties of $\circledast$ are also available. These are given in Sections 2.9 and 2.10.

One of the important features of the partitioned Kronecker product, which is applicable in the differentiation of partitioned matrix functions, is that it helps us to transform partitioned matrices into column vectors. We discuss this in the next section. Some statistical applications of $\circledast$ are discussed in Chapter IV.

Definition 2.7.2 (Khatri & Rao (1968)). Let

$$A = (A_{.1}, \ldots, A_{.q}): m \times q; \quad B = (B_{.1}, \ldots, B_{.q}): n \times q$$

be partitioned matrices, then $\odot$ is defined as

$$A \odot B = (A_{.1} \otimes B_{.1} \quad A_{.2} \otimes B_{.2} \quad \ldots \quad A_{.q} \otimes B_{.q}).$$

Definition 2.7.3 (Khatri (1971)). Let $A: m \times n$, $B: p \times q$ be matrices which are partitioned as

$$A = \overset{\begin{matrix} n_1 & n_2 & \ldots & n_r \end{matrix}}{[A^1 \vdots A^2 \vdots \ \ldots \ \vdots A^r]}, \quad B = \overset{\begin{matrix} q_1 & q_2 & \ldots & q_r \end{matrix}}{[B^1 \vdots B^2 \vdots \ \ldots \ \vdots B^r]};$$

then the matrix product ⓔ is defined as

$$A \text{ ⓔ } B = [A^1 \otimes B^1 \vdots A^2 \otimes B^2 \vdots \ \ldots \ \vdots A^r \otimes B^r].$$

It is easily seen that the matrix product ⓔ is an extension of $\odot$.

The matrix product $A \odot B$ (Khatri & Rao (1968)) and A ⓔ B (Khatri (1971)), useful in solving certain functional equations needed in characterizing multivariate distributions, may be obtained from very specialized cases of the matrix product A ⓟ B.

If no row-wise partitioning is done, and column-wise partitioning is done such that the number of column blocks in A and B is the same, say r in each case, then we have

$$A \text{ ⓟ } B = [A^j \text{ (\#) } B], \quad j = 1, 2, \ldots, r,$$

where

$$A^j \text{ (\#) } B = [A^j \otimes B^1 \vdots \ \ldots \ \vdots A^j \otimes B^r].$$

If we now retain only $A^j \otimes B^j$, $j = 1, 2, \ldots, r$, the 1 mod(r+1)-th of the r^2 column blocks in A ⓟ B, we obtain A ⓔ B. If further,

$k_j = t_v = 1$, i.e., column-wise partitioning in A and B is done by single columns, then A ⓔ B becomes A ⊙ B.

2.8 Extensions of Auxiliary Matrices

Here our main attempt is to extend the concepts $I_{(m)}$ and $I_{(n)}$ for non-partitioned identity matrices mentioned in Section 2.4 to partitioned identity matrices. Since the sums of partitioned matrices are obtained from the rule of row by row sum (or column by column sum) and products from row by column multiplication, certain extensions of ordinary matrices to partitioned matrices are easily available by treating matrices as elements.

Definition 2.8.1 Let Y be partitioned into

$$Y = \begin{bmatrix} Y^{11} & \cdots & Y^{1n} \\ \vdots & & \\ Y^{m1} & \cdots & Y^{mn} \end{bmatrix}, \qquad (2.8.1)$$

where Y^{ij} is the (i,j)-th submatrix of Y. Then $[Y^{i1} \; \dots \; Y^{in}]$

and $\begin{bmatrix} Y^{1j} \\ \vdots \\ Y^{mj} \end{bmatrix}$ are the i-th row block and j-th column block of

Y respectively.

Definition 2.8.2 Let $I:pq\times pq$ be partitioned into mn row blocks and mn column blocks according to the scheme

$p_1q_1 \dots p_mq_1 \dots \quad \dots p_1q_n \dots p_mq_n$ as

$$I:pq\times pq = \begin{array}{c} \\ p_1q_1 \\ \vdots \\ p_mq_1 \\ \vdots \\ \vdots \\ p_1q_n \\ \vdots \\ p_mq_n \end{array} \begin{array}{c} p_1q_1 \cdots p_mq_1 \cdots \quad \cdots p_1q_n \cdots p_mq_n \\ \left[\begin{array}{c|c|ccc|c} I & 0 & 0 & & & 0 \\ \hline 0 & I & 0 & & & 0 \\ \hline 0 & 0 & \ddots & & & \\ & & & \ddots & & \\ & & & & \ddots & \\ & & & & & 0 \\ \hline 0 & 0 & & & 0 & I \end{array}\right] \end{array} \qquad (2.8.2)$$

such that $\sum_{1}^{m} p_i = p$ and $\sum_{1}^{n} q_j = q$. Suppose we rearrange the row blocks of (2.8.2) by taking every m-th row block starting with the first, then every m-th row block starting with the second etc. We denote this rearranged matrix by $_{[m]}I$.

<u>Definition 2.8.3</u> Suppose that $I:pq\times pq$ is partitioned into row and column blocks according to the scheme $p_1q_1 \cdots p_1q_n \cdots \quad \cdots p_mq_1 \cdots p_mq_n$. Then we define a matrix $I_{[n]}$ which is obtained on rearranging the row blocks of this I by taking every n-th row block starting with the first, and then every n-th row block starting with the second, etc.

<u>Example 2.8.1</u> Let $m = 2$, $n = 3$; $p_1 = 2$, $p_2 = 1$, $q_1 = 2$, $q_2 = 1$, $q_3 = 1$. If

$$I{:}12\times12 = \begin{array}{c} \\ 4\\2\\2\\1\\2\\1 \end{array}\begin{array}{c} \begin{array}{cccccc} 4 & 2 & 2 & 1 & 2 & 1 \end{array} \\ \begin{bmatrix} I_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

then

$$[2]^{I} = \begin{array}{c} \\ 4\\2\\2\\2\\1\\1 \end{array}\begin{array}{c} \begin{array}{cccccc} 4 & 2 & 2 & 1 & 2 & 1 \end{array} \\ \begin{bmatrix} I_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_2 & 0 \\ 0 & I_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} .$$

Similarly, if

$$I{:}12\times12 = \begin{array}{c} \\ 4\\2\\2\\2\\1\\1 \end{array}\begin{array}{c} \begin{array}{cccccc} 4 & 2 & 2 & 2 & 1 & 1 \end{array} \\ \begin{bmatrix} I_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

then

$$I_{[3]} = \begin{array}{c} \\ 4 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} 4 & 2 & 2 & 2 & 1 & 1 \end{array} \\ \begin{bmatrix} I_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_2 & 0 & 0 \\ 0 & I_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & I_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} .$$

One may observe that in the definitions of ${}_{[m]}I$ and $I_{[n]}$ we have rearranged the row blocks of suitably partitioned $I{:}pq\times pq$; whereas in defining $I_{(m)}$ and $I_{(n)}$, the rows of non-partitioned $I{:}mn\times mn$ are rearranged. Thus we may say that partitioned matrices ${}_{[m]}I$ and $I_{[n]}$ are block generalizations of the auxiliary matrices $I_{(m)}$ and $I_{(n)}$. It would have been more appropriate to denote by $I_{[m]}$ and $I_{[n]}$ the matrices introduced in Definitions 2.8.2 and 2.8.3, as extensions of $I_{(m)}$ and $I_{(n)}$ respectively, but there are certain difficulties. Tracy & Dwyer (1969) were able to denote the auxiliary matrices by $I_{(m)}$ and $I_{(n)}$, since when $m = n$, $I_{(m)} = I_{(n)}$; whereas in our case $I_{[m]} \neq I_{[n]}$ for $m = n$. We illustrate this by the following example:

<u>Example 2.8.2</u> Let $m = 2$, $n = 2$; $p_1 = 2$, $p_2 = 1$, $q_1 = 1, q_2 = 2$.

$$I:9\times 9 = \begin{array}{c} \\ 2 \\ 1 \\ 4 \\ 2 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 1 & 4 & 2 \end{array} \\ \begin{bmatrix} I_2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & I_4 & 0 \\ 0 & 0 & 0 & I_2 \end{bmatrix} \end{array},$$

then

$$[m]^{I}=[2]^{I} = \begin{array}{c} \\ 2 \\ 4 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 1 & 4 & 2 \end{array} \\ \begin{bmatrix} I_2 & 0 & 0 & 0 \\ 0 & 0 & I_4 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & I_2 \end{bmatrix} \end{array}$$

and if

$$I:9\times 9 = \begin{array}{c} \\ 2 \\ 4 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 4 & 1 & 2 \end{array} \\ \begin{bmatrix} I_2 & 0 & 0 & 0 \\ 0 & I_4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

then

$$I_{[n]}=I_{[2]} = \begin{array}{c} \\ 2 \\ 1 \\ 4 \\ 2 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 4 & 1 & 2 \end{array} \\ \begin{bmatrix} I_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & I_4 & 0 & 0 \\ 0 & 0 & 0 & I_2 \end{bmatrix} \end{array}.$$

Here easily we see that ${}_{[m]}I \neq I_{[n]}$ for $m = n = 2$. Such an ambiguity will not arise if the rearrangements of row blocks of the partitioned identity matrix I are denoted by ${}_{[m]}I$ and $I_{[n]}$ when row and column partitioning schemes of I are $p_1q_1 \cdots p_mq_1 \cdots \quad \cdots p_1q_n \cdots p_mq_n$ and $p_1q_1 \cdots p_1q_n \cdots \quad \cdots p_mq_1 \cdots p_mq_n$ respectively. Since ${}_{[m]}I$ and $I_{[n]}$ are notationally different, no relationship between them is necessary.

These extended auxiliary matrices are used to rearrange the rows of an arbitrarily partitioned matrix. Suppose $P{:}pq\times s$ be partitioned into mn row blocks and into any arbitrary number of column blocks. Then ${}_{[m]}P$ and $P_{[n]}$ are rearrangements of row blocks of P by taking every m-th (n-th) row block starting with the first, then every m-th (n-th) row block starting with the second, etc., when the row partitioning schemes are $p_1q_1 \ldots p_mq_1 \ldots \ldots p_1q_n \ldots p_mq_n$ and $p_1q_1 \ldots p_1q_n \ldots \ldots p_mq_1 \ldots p_mq_n$ respectively. In such a case we see that ${}_{[m]}IP = {}_{[m]}P$ and $(I_{[n]})P = P_{[n]}$. We illustrate these in the following example.

Example 2.8.2 Let m = 3, n = 2; i,i' = 1,...,6; j = 1, 2; j' = 1, 2, 3; and $P{:}pq\times s$ a partitioned matrix. Then

$$P = \begin{matrix} & s_1 & s_2 \\ p_1q_1 & M^{11} & M^{12} \\ p_2q_1 & M^{21} & M^{22} \\ p_3q_1 & M^{31} & M^{32} \\ p_1q_2 & M^{41} & M^{42} \\ p_2q_2 & M^{51} & M^{52} \\ p_3q_2 & M^{61} & M^{62} \end{matrix} \Longrightarrow {}_{[3]}P = \begin{matrix} & s_1 & s_2 \\ p_1q_1 & M^{11} & M^{12} \\ p_1q_2 & M^{41} & M^{42} \\ p_2q_1 & M^{21} & M^{22} \\ p_2q_2 & M^{51} & M^{52} \\ p_3q_1 & M^{31} & M^{32} \\ p_3q_2 & M^{61} & M^{62} \end{matrix},$$

and

$$P = \begin{matrix} & s_1' & s_2' & s_3' \\ p_1q_1 & M_{11} & M_{12} & M_{13} \\ p_1q_2 & M_{21} & M_{22} & M_{23} \\ p_2q_1 & M_{31} & M_{32} & M_{33} \\ p_2q_2 & M_{41} & M_{42} & M_{43} \\ p_3q_1 & M_{51} & M_{52} & M_{53} \\ p_3q_2 & M_{61} & M_{62} & M_{63} \end{matrix} \Longrightarrow P_{[2]} = \begin{matrix} & s_1' & s_2' & s_3' \\ p_1q_1 & M_{11} & M_{12} & M_{13} \\ p_2q_1 & M_{31} & M_{32} & M_{33} \\ p_3q_1 & M_{51} & M_{52} & M_{53} \\ p_1q_2 & M_{21} & M_{22} & M_{23} \\ p_2q_2 & M_{41} & M_{42} & M_{43} \\ p_3q_2 & M_{61} & M_{62} & M_{63} \end{matrix} .$$

Also we can see that ${}_{[3]}IP = {}_{[3]}P$ and $(I_{[2]})P = P_{[2]}$.

We use the following notation:

$$\mathrm{Diag}[A_{11} \quad A_{22} \quad \cdots \quad A_{mm}] = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & & \\ \vdots & & \ddots & \\ 0 & \cdots & & A_{mm} \end{bmatrix} .$$

Two additional partitioned auxiliary matrices are defined as follows:

<u>Definition 2.8.4</u> Let $I{:}pq \times pq$ be partitioned as in (2.8.2), where $p = \sum_{i=1}^{m} p_i$ and $q = \sum_{j=1}^{n} q_j$. Then by ${}_{<m>}I$ we mean an auxiliary partitioned matrix given by

$$\begin{matrix} & p_1q_1 & \cdots & p_mq_1 & \cdots & \cdots & p_1q_n & \cdots & p_mq_n \\ {}_{<m>}I = \mathrm{Diag}[& I_{(p_1)} & \cdots & I_{(p_m)} & \cdots & \cdots & I_{(p_1)} & \cdots & I_{(p_m)}] \end{matrix} \quad (2.8.3)$$

where the blocks $I_{(p_i)}$, $i = 1,\ldots,m$, are as in Definition 2.4.1.

<u>Definition 2.8.5</u> If the scheme of partitioning of $I:pq\times pq$ is as in Definition 2.8.3, then we define an auxiliary partitioned matrix

$$I_{<n>} = \underset{}{\mathrm{Diag}}[\overset{p_1q_1}{I_{(q_1)}} \cdots \overset{p_1q_n}{I_{(q_n)}} \cdots \cdots \overset{p_mq_1}{I_{(q_1)}} \cdots \overset{p_mq_n}{I_{(q_n)}}] \quad (2.8.4)$$

where the $I_{(q_j)}$, $j = 1,\ldots,n$, are as in Definition 2.4.1 and

$\sum_{i=1}^{m} p_i = p, \quad \sum_{j=1}^{n} q_j = q$.

<u>Example 2.8.3</u> Let $m = 3$, $n = 2$ so that $p_1+p_2+p_3 = p$ and $q_1+q_2 = q$. If

$$I:pq\times pq = \mathrm{Diag}[I_{p_1q_1} \; I_{p_2q_1} \; I_{p_3q_1} \; I_{p_1q_2} \; I_{p_2q_2} \; I_{p_3q_2}]$$

then

$$_{<3>}I = \begin{array}{c} \\ p_1q_1 \\ p_2q_1 \\ p_3q_1 \\ p_1q_2 \\ p_2q_2 \\ p_3q_2 \end{array} \begin{array}{cccccc} p_1q_1 & p_2q_1 & p_3q_1 & p_1q_2 & p_2q_2 & p_3q_2 \\ \left[I_{(p_1)} \right. & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{(p_2)} & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{(p_3)} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{(p_1)} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{(p_2)} & 0 \\ 0 & 0 & 0 & 0 & 0 & \left. I_{(p_3)} \right] \end{array} .$$

Similarly one can see that

$$I_{<2>} = \mathrm{Diag}[\overset{p_1q_1}{I_{(q_1)}} \; \overset{p_1q_2}{I_{(q_2)}} \; \overset{p_2q_1}{I_{(q_1)}} \; \overset{p_2q_2}{I_{(q_2)}} \; \overset{p_3q_1}{I_{(q_1)}} \; \overset{p_3q_2}{I_{(q_2)}}] .$$

Here again we note that ${}_{<m>}I \neq I_{<n>}$ for m = n.

With the help of the auxiliary partitioned matrices ${}_{[m]}I$ ${}_{<m>}I$, $I_{[n]}$ and $I_{<n>}$, we have the following definition:

Definition 2.8.6 For a partitioned matrix I:pq×pq, let ${}_{[m]}I$, ${}_{<m>}I$, $I_{[n]}$ and $I_{<n>}$ be as in Definitions 2.8.2 - 2.8.5. Then we define auxiliary matrices ${}_{\{m\}}I$ and $I_{\{n\}}$ so that

$$({}_{[m]}I)({}_{<m>}I) = {}_{\{m\}}I \tag{2.8.5}$$

$$(I_{[n]})(I_{<n>}) = I_{\{n\}} \ . \tag{2.8.6}$$

The matrices ${}_{\{m\}}I$ and $I_{\{n\}}$ are used to perform certain operations on suitably partitioned matrices. For non-partitioned matrices, these operations are carried out with the help of the auxiliary matrices $I_{(m)}$ and $I_{(n)}$.

Example 2.8.4 For the matrices used in Example 2.8.3, we have

$$ {}_{\{3\}}I = \begin{bmatrix} I_{(p_1)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{(p_1)} & 0 & 0 \\ 0 & I_{(p_2)} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{(p_2)} & 0 \\ 0 & 0 & I_{(p_3)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{(p_3)} \end{bmatrix} $$

and

$$I_{\{2\}} = \begin{bmatrix} I_{(q_1)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{(q_1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{(q_1)} & 0 \\ 0 & I_{(q_2)} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{(q_2)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{(q_2)} \end{bmatrix}$$

Also it may be verified that ${I_{\{2\}}}^T = {}_{\{3\}}I$, $(I_{\{2\}})({}_{\{3\}}I) = ({}_{\{3\}}I)(I_{\{2\}}) = I$, and $I_{\{2\}} \neq {}_{\{3\}}I$.

Some important properties of ${}_{\{m\}}I$, $I_{\{n\}}$ are stated in the following theorem:

Theorem 2.8.1 For auxiliary partitioned matrices ${}_{\{m\}}I$ and $I_{\{n\}}$, we have

(i) ${}_{\{m\}}I$ and $I_{\{n\}}$ are conformable for partitioned matrix multiplication (2.8.7)

(ii) $({}_{\{m\}}I)(I_{\{n\}}) = (I_{\{n\}})({}_{\{m\}}I) = I$ (2.8.8)

(iii) ${}_{\{m\}}I^T = I_{\{n\}}$ (2.8.9)

(iv) ${}_{\{m\}}I \neq I_{\{n\}}$ even if $m = n$ (2.8.10)

(v) ${}_{\{m\}}I + I_{\{n\}}$ is a symmetric partitioned matrix (2.8.11)

(vi) $|{}_{\{m\}}I| = |I_{\{n\}}| = \pm 1$. (2.8.12)

Proof: Since the row-wise partitioning scheme of after-factor is the same as the column-wise partitioning scheme of fore-factor hence both $({}_{\{m\}}I)(I_{\{n\}})$ and $(I_{\{n\}})({}_{\{m\}}I)$ exist, which proves (i). Other results have a straightforward proof. A very simple proof of (ii) is given in Section 2.10.

Thus we see that ${}_{\{m\}}I$ and $I_{\{n\}}$ are certain extensions of $I_{(m)}$ and $I_{(n)}$ respectively. These extended matrices are useful in carrying out certain between and within block rearrangements. Such rearrangements are required in developing a partitioned matrix differential calculus. Some new properties of the matrix product $\textcircled{\pi}$, which involve auxiliary partitioned matrices, are also available.

2.9 Additional Properties of the Matrix Product $\textcircled{\pi}$

The following theorem is a straightforward extension of Theorem 2.5.1.

Theorem 2.9.1 Let A be partitioned into m row blocks and n column blocks, and B be partitioned into g row blocks and h column blocks. Then the following equalities hold:

(i) $${}_{\{g\}}I(A^T \textcircled{\pi} B) = (B \textcircled{\pi} A^T)I_{\{h\}} \tag{2.9.1}$$

(ii) $${}_{\{m\}}I(B \textcircled{\pi} A) = (A \textcircled{\pi} B)I_{\{n\}} \tag{2.9.2}$$

(iii) $$A \textcircled{\pi} B = {}_{\{m\}}I(B \textcircled{\pi} A)I_{\{h\}} \tag{2.9.3}$$

These results are easily verified.

Theorem 2.9.1 is useful in simplifying partitioned matrix differentiation.

2.10 Extension of the Column Vectors Y_r and Y_c

Sometimes it is convenient to extend the use of the fundamental procedure of column vector representation discussed in Section 2.6 to the case where a vector is regarded as consisting of smaller subvectors and a matrix consists of two or more submatrices in a partitioned form. In our discussion of partitioned matrices, here and also in sections below, we assume that the matrices involved are partitioned conformably both for addition and for multiplication. However, when we come across non-linear partitioned matrix functions such as $Y = XX^T$, $Y = XX^TX$ etc., then the conformability is obvious and the desired results concerning these matrices are very easily obtained.

Definition 2.10.1 Let a matrix $Y{:}p\times q$ be partitioned as

$$Y = p_i \overset{q_j}{\left[Y^{ij}\right]},$$

where $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$; and $\sum_{i=1}^{m} p_i = p$, $\sum_{j=1}^{n} q_j = q$. Then we define column vector representations Y_R and Y_C of Y as below:

$$Y_R = \begin{matrix} p_1q_1 \\ \vdots \\ p_1q_n \\ \vdots \\ \vdots \\ p_mq_1 \\ \vdots \\ p_mq_n \end{matrix} \begin{bmatrix} Y^{11}{}_r \\ \vdots \\ Y^{1n}{}_r \\ \vdots \\ \vdots \\ Y^{m1}{}_r \\ \vdots \\ Y^{mn}{}_r \end{bmatrix} \quad \text{and} \quad Y_C = \begin{matrix} p_1q_1 \\ \vdots \\ p_mq_1 \\ \vdots \\ \vdots \\ p_1q_n \\ \vdots \\ p_mq_n \end{matrix} \begin{bmatrix} Y^{11}{}_c \\ \vdots \\ Y^{m1}{}_c \\ \vdots \\ \vdots \\ Y^{1n}{}_c \\ \vdots \\ Y^{mn}{}_c \end{bmatrix} .$$

In fact, Y_R and Y_C are block generalizations of Y_r and Y_c, respectively, for a partitioned matrix Y.

<u>Example 2.10.1</u> Let

$$Y = \begin{matrix} & q_1 & q_2 \\ p_1 \\ p_2 \\ p_3 \end{matrix} \begin{bmatrix} Y^{11} & Y^{12} \\ Y^{21} & Y^{22} \\ Y^{31} & Y^{32} \end{bmatrix} = \left[\begin{array}{c|cc} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \hline y_{31} & y_{32} & y_{33} \\ \hline y_{41} & y_{42} & y_{43} \\ y_{51} & y_{52} & y_{53} \end{array}\right] ,$$

where $m = 3$, $n = 2$, and $p_1 = 2$, $p_2 = 1$, $p_3 = 2$; $q_1 = 1$, $q_2 = 2$.

Then

$$Y_R = \begin{array}{c|c|}
 & 1 \\ \hline
2 & y_{11} \\
 & y_{21} \\ \hline
 & y_{12} \\
4 & y_{13} \\
 & y_{22} \\
 & y_{23} \\ \hline
1 & y_{31} \\ \hline
2 & y_{32} \\
 & y_{33} \\ \hline
2 & y_{41} \\
 & y_{51} \\ \hline
 & y_{42} \\
4 & y_{43} \\
 & y_{52} \\
 & y_{53} \\ \hline
\end{array} \quad \text{and} \quad Y_C = \begin{array}{c|c|}
 & 1 \\ \hline
2 & y_{11} \\
 & y_{21} \\ \hline
1 & y_{31} \\ \hline
2 & y_{41} \\
 & y_{51} \\ \hline
 & y_{12} \\
4 & y_{22} \\
 & y_{13} \\
 & y_{23} \\ \hline
2 & y_{32} \\
 & y_{33} \\ \hline
 & y_{42} \\
4 & y_{52} \\
 & y_{43} \\
 & y_{53} \\ \hline
\end{array} .$$

If, for all i and j, $p_i = q_j = 1$, then we observe that $Y_R = Y_r$ and $Y_C = Y_c$.

Now we apply the auxiliary matrices ${}_{\{m\}}I$ and $I_{\{n\}}$ to establish certain interrelationships between Y_R and Y_C which are given below:

<u>Theorem 2.10.1</u> For a partitioned matrix $Y{:}p{\times}q$ considered in

Definition 2.10.1, we have

(i) $Y^T{}_C = Y_R$ (2.10.1)

(ii) $Y_R = {}_{\{m\}}IY_C$ (2.10.2)

(iii) $Y_C = (I_{\{n\}})Y_R$ (2.10.3)

These results are extensions of (2.6.3.9), (2.6.2.1) and (2.6.2.2) respectively.

Proof: (i) Obvious.

(ii) Using Definitions 2.8.2, 2.8.4, 2.10.1 and equation (2.8.5), we obtain

$$
{}_{\{m\}}IY_C = {}_{[m]}I \begin{bmatrix} Y^{11}{}_r \\ \cdot \\ \cdot \\ \cdot \\ Y^{m1}{}_r \\ \cdot \\ \cdot \\ \cdot \\ \\ \cdot \\ \cdot \\ \cdot \\ Y^{1n}{}_r \\ \cdot \\ \cdot \\ \cdot \\ Y^{mn}{}_r \end{bmatrix} = \begin{bmatrix} Y^{11}{}_r \\ \cdot \\ \cdot \\ \cdot \\ Y^{1n}{}_r \\ \cdot \\ \cdot \\ \cdot \\ \\ \cdot \\ \cdot \\ \cdot \\ Y^{m1}{}_r \\ \cdot \\ \cdot \\ \cdot \\ Y^{mn}{}_r \end{bmatrix} = Y_R \quad .
$$

(iii) Follows from (ii).

As a consequence of (2.10.2) and (2.10.3), we get

$Y_R = {}_{\{m\}}I(I_{\{n\}})Y_R$

and

$$Y_C = (I_{\{n\}})({}_{\{m\}}I)Y_C \quad .$$

From these it is clear that $(I_{\{n\}})({}_{\{m\}}I) = ({}_{\{m\}}I)(I_{\{n\}}) = I$, which proves (2.8.8).

Thus we observe that ${}_{\{m\}}I$, $I_{\{n\}}$ are used to perform certain extended operations.

Example 2.10.2 For the matrices given in Examples 2.8.4 and 2.10.1, one can verify that

$$Y_R = {}_{\{3\}}IY_C \tag{2.10.4}$$

and

$$Y_C = (I_{\{2\}})Y_R \ . \tag{2.10.5}$$

Expressions (2.10.4) and (2.10.5) may be compared with $Y_r = I_{(p)}Y_c$ and $Y_c = I_{(q)}Y_r$ respectively, where $p_1+p_2+p_3 = p$, $q_1+q_2 = q$.

We need the column vector representation of a partitioned matrix function as a function of the column vector representation of a partitioned matrix variable. This is achieved by the following basic theorem:

Theorem 2.10.2 For any partitioned matrix function $Y = AXB$, we have

$$Y_R = (A \circledast B^T)X_R \tag{2.10.6}$$

and

$$Y_C = (B^T \circledast A)X_C \tag{2.10.7}$$

Proof: It can be seen that

$$Y^{ij}{}_r = \left[A^{i\cdot} \circledast B^{\cdot j^T}\right] X_R$$

is a typical subvector of Y_R . Suitable rearrangement of these subvectors yields (2.10.6).

Similarly, (2.10.7) can be proved.

<u>Corollary 2.10.1</u> Let $Y = AX^TB$ where X is partitioned into m row blocks and n column blocks. Then

$$Y_R = (A \textcircled{\pi} B^T)(I_{\{n\}})X_R \qquad (2.10.8)$$

and

$$Y_C = (B^T \textcircled{\pi} A)({}_{\{m\}}I)X_C \ . \qquad (2.10.9)$$

When the matrices involved are non-partitioned, then the above results may be compared with (2.6.3.11) - (2.6.3.14).

Theorem 2.10.2 and Corollary 2.10.1 may be extended to more general partitioned matrix functions, for example, $Y = AXBX^TC$, by using property (2.7.6) and the method of induction. These extended results are used in partitioned matrix differentiation.

2.11 Two Extensions of Khatri's Product and Their Properties

Let

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & \\ \vdots & & & \\ A_{m1} & & \cdots & A_{mn} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & & & \\ \vdots & & & \\ B_{p1} & & \cdots & B_{pn} \end{bmatrix} \tag{2.11.1}$$

be partitioned matrices. Then we define the following two extensions Ⓔ and ⊛ of the matrix product ⓔ defined by Khatri (1971):

Definition 2.11.1

$$A \text{Ⓔ} B = \begin{bmatrix} A_{11}\otimes B_{11} & A_{12}\otimes B_{12} & \cdots & A_{1n}\otimes B_{1n} \\ \vdots & \vdots & & \vdots \\ A_{11}\otimes B_{p1} & A_{12}\otimes B_{p2} & & A_{1n}\otimes B_{pn} \\ \vdots & \vdots & & \vdots \\ A_{m1}\otimes B_{11} & A_{m2}\otimes B_{12} & & A_{mn}\otimes B_{1n} \\ \vdots & \vdots & & \vdots \\ A_{m1}\otimes B_{p1} & A_{m2}\otimes B_{p2} & & A_{mn}\otimes B_{pn} \end{bmatrix} \tag{2.11.2}$$

Definition 2.11.2

$$A \circledast B = \begin{bmatrix} A_{11}\otimes B_{11} & A_{12}\otimes B_{12} & \cdots & A_{1n}\otimes B_{1n} \\ A_{21}\otimes B_{21} & A_{22}\otimes B_{22} & \cdots & A_{2n}\otimes B_{2n} \\ \vdots & & & \\ A_{m1}\otimes B_{m1} & A_{m2}\otimes B_{m2} & \cdots & A_{mn}\otimes B_{mn} \end{bmatrix} \tag{2.11.3}$$

$$= [A_{ij}\otimes B_{ij}] \tag{2.11.4}$$

where p = m in (2.11.1).

Below we assume that the matrices involved are conformable for matrix products and other matrix operations under consideration.

Certain properties of Ⓔ are the following:

(i) $A \circledE B$ and $B \circledE A$ are coexistent (2.11.5)

(ii) $A \circledE B \neq B \circledE A$ in general (2.11.6)

(iii) $A \circledE 0 = 0 \circledE A = 0$ (2.11.7)

(iv) $\alpha A \circledE \beta B = \alpha\beta(A \circledE B)$, for scalars α, β (2.11.8)

(v) $(A+B) \circledE (C+D) = A \circledE C + A \circledE D + B \circledE C + B \circledE D$ (2.11.9)

(vi) If A and B are as in (2.11.1), then

$$A \circledE B = {}_{\{m\}}I(B \circledE A), \tag{2.11.10}$$

where ${}_{\{m\}}I$ is as in Section 2.8.

Some of the properties of Ⓔ are extensions of the corresponding properties of ⓔ, (see Khatri, 1971, p. 76, (iv) and (v)). These are:

If

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & & \\ C_{q1} & \cdots & C_{qn} \end{bmatrix} \tag{2.11.11}$$

then

$$(A \circledE B) \circledE C = A \circledE (B \circledE C) \tag{2.11.12}$$

If

$$G = \begin{bmatrix} G_{11} & \cdots & G_{1m} \\ \vdots & & \\ G_{k1} & \cdots & G_{km} \end{bmatrix}, \quad H = \begin{bmatrix} H_{11} & \cdots & H_{1p} \\ \vdots & & \\ H_{\ell 1} & & H_{\ell p} \end{bmatrix} \tag{2.11.13}$$

then

$$(G \textcircled{\pi} H)(A \textcircled{E} B) = GA \textcircled{E} HB \qquad (2.11.14)$$

where $\textcircled{\pi}$ is the partitioned Kronecker product defined in Section 2.7. Khatri (1971) studied various other properties of $\textcircled{e}$ with the help of the Kronecker product $\otimes$. Since $\textcircled{\pi}$ and $\textcircled{E}$ are extensions of $\otimes$ and $\textcircled{e}$ respectively, we may as well investigate additional properties of $\textcircled{E}$ with the help of $\textcircled{\pi}$ and some other related concepts.

The matrix product $\circledast$ defined in (2.11.3) possesses the following properties:

(i) $A \circledast B$ and $B \circledast A$ are coexistent (2.11.15)

(ii) $A \circledast B \neq B \circledast A$ in general (2.11.16)

(iii) $A \circledast 0 = 0 \circledast A = 0$ (2.11.17)

(iv) $\alpha A \circledast \beta B = \alpha\beta(A \circledast B)$, for scalars α, β (2.11.18)

(v) $(A+B) \circledast (C+D) = A \circledast C + A \circledast D + B \circledast C + B \circledast D$ (2.11.19)

(vi) $(A \circledast B) \circledast C = A \circledast (B \circledast C)$ (2.11.20)

(vii) $(A \circledast B)^T = A^T \circledast B^T$ (2.11.21)

Definition 2.11.3 (Generalized Schur Product). If $A = [A_{ij}]$, $B = [B_{ij}]$ are partitioned matrices such that $A_{ij}B_{ij}$ exists for all i, j, then a generalization * of the Schur product ×× is given by

$$A*B = [A_{ij}B_{ij}], \qquad (2.11.22)$$

where $A_{ij}B_{ij}$ is the (i,j)-th partition of A*B.

Rao & Mitra (1971) introduced the matrix product * defined as in (2.11.22). However, they did not mention any of its properties. It may be verified that all the properties of the

Schur product ×× are not extendable to the <u>generalized Schur product</u>. Some properties of the Schur product are given in Rao & Mitra (1971, pp. 11-12). A few properties of the generalized Schur product are:

(i) Existence of A*B does not imply the existence of B*A in general (2.11.23)

(ii) $A*O = O*A = O$ (2.11.24)

(iii) $A*B \neq B*A$ in general (2.11.25)

(iv) $(A*B)*C = A*(B*C)$ (2.11.26)

(v) $\alpha A*\beta B = \alpha\beta(A*B)$, α,β are scalars (2.11.27)

(vi) $(A+B)*(C+D) = A*C + A*D + B*C + B*D$ (2.11.28)

(vii) A very interesting relationship between the matrix products * defined by Rao & Mitra (1971) and ⊛ introduced in this section is the following:

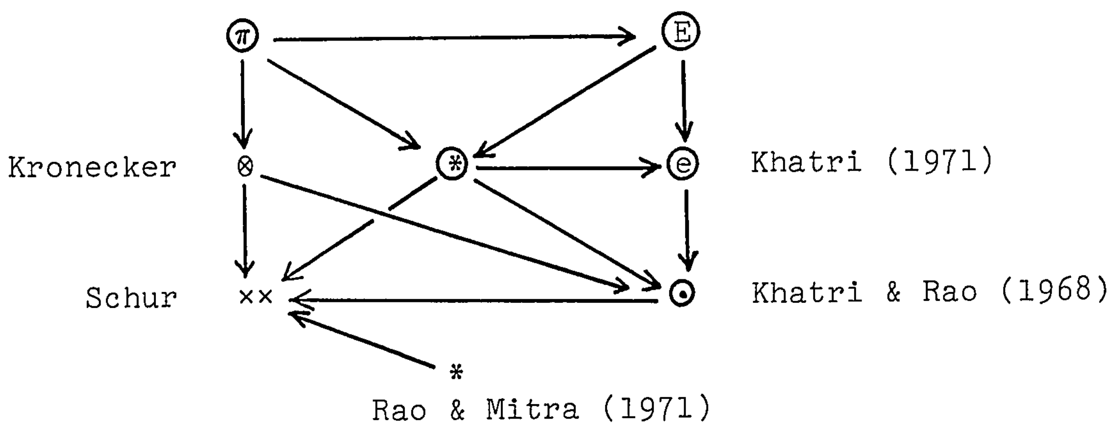
$$(A*B) \circledast (C*D) = (A \circledast C)*(B \circledast D) \qquad (2.11.29)$$

(viii) $$tr(A*B*C) = \sum_i (A_{ii})_r^T (I \otimes B_{ii})(C_{ii}^T)_r \qquad (2.11.30)$$

(ix) $$(A*B)^T = B^T*A^T \qquad (2.11.31)$$

Various specializations concerning some matrix products discussed in this work are demonstrated by the following diagram:

FIGURE 2.11.1 SPECIAL MATRIX PRODUCTS



CHAPTER III

DIFFERENTIATION OF FUNCTIONS OF MATRICES

3.1 Introduction

The problem of matrix differentiation for the purpose of statistical application has been considered by Dwyer & MacPhail (1948), Coy (1955), Wrobleski (1963), Dwyer (1967), Neudecker (1967, 1968, 1969b) and Tracy & Dwyer (1969). A procedure different from that of the above authors but valid for differentiating functions of square matrices only is found in the papers by Aitken (1953), Capelli (1887), Garding (1948), Turnbull (1927-29, 1930-31a, 1930-31b, 1948) and Wallace (1953). The latter method is based on the Cayleyan operator and has found some applications in multivariate analysis, as pointed out by Gordon (1967) and Stroud (1971).

The matrix calculus presented by Dwyer (1967) is general enough to include the results of previous authors as particular cases, both in the theory and in applications. In later papers, Neudecker (1969b) and Tracy & Dwyer (1969) extended the concept of this matrix calculus to obtain first and second order matrix derivatives of matrix functions. Papers by Dwyer (1967) and Tracy & Dwyer (1969) were based on the first order matrix derivatives of J-type and K-type matrices as defined by Dwyer & MacPhail (1948). These matrices are like E_{ij} for X and Y respectively. In the paper by Tracy & Dwyer (1969), expressions for $\frac{\partial Y_r}{\partial X_r}$,, $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$, and $\frac{\partial Y_c}{\partial X_c}$, when the matrix Y is some commonly occurring matrix function of a matrix variable X , are presented by making use of auxiliary matrices $I_{(k)}$. A new procedure of identifying matrices of partial derivatives of some matrix functions involving (i) ordinary matrix products, (ii) Kronecker matrix products, without using the auxiliary matrix $I_{(k)}$ and matrix derivatives of the J-type and K-type was suggested by Neudecker (1969b). However, in his consideration, auxiliary matrices $I_{(k)}$ for various positive integers k would have simplified the matrix derivative results to a great extent. In the above papers the authors considered some particular matrix functions and obtained their matrix derivatives.

Hausdorff differentiability of a matrix function of a square matrix variable has been studied by Rinehart (1966a, 1966b) and Powers (1971). Powers gave a new diagonalizing matrix which, together with the technique in Neudecker (1969b), provides a simpler computation of Hausdorff derivatives than that given by Rinehart (1966a). According to Rinehart (1966a) and Powers (1971), the differential df(X) of a matrix function f(X) of a square matrix variable X may be expressed in terms of differentials dP and $d\Lambda$, where $P^{-1}XP = \Lambda$ and Λ is a diagonal matrix.

In this chapter we are concerned with the differential calculus of matrix functions involving ordinary matrix products some of which are more general than those considered in Section 2.3. This approach generalizes some of the theoretical results on matrix differentiation given by Dwyer (1967), Neudecker (1969b) and Tracy & Dwyer (1969). Besides the above generalizations, introduced and studied in this chapter are some new concepts such as matrix derivative transformations, mixed partial matrix derivatives, partitioned matrix differentiation and differentiation of functions of matrices involving equality relationships among their scalar elements. Certain basic facts given by Dwyer (1970) concerning matrix derivatives are verified.

Some further results based on the above concepts are also presented. In this presentation, the J-type and K-type matrix derivatives are not used at any stage because the fundamental concepts of $\frac{\partial X}{\partial \langle X \rangle}$, $\frac{\partial \langle X \rangle}{\partial X}$, $\frac{\partial Y}{\partial \langle X \rangle}$ and $\frac{\partial \langle Y \rangle}{\partial X}$ are not used in this thesis and only their transforms to vectors such as $\frac{\partial Y_r}{\partial X_r}$ are treated. However, Sections 2.4-2.10 of the previous chapter, in coordination with the procedure available in Neudecker (1969b) and Tracy & Dwyer (1969), play a major role. This theory is developed in the field of real numbers over which all the vector spaces under consideration are defined.

An explanation of the term "matrix element" as opposed to "scalar element" is also very essential in this chapter. Throughout our discussion, for any matrix X, we denote its (i,j)-th <u>matrix element</u> by $\langle x_{ij} \rangle$, which is (i,j,x_{ij}) in the terminology of Dwyer (1970, p. 5), and its (i,j)-th <u>scalar element</u> by x_{ij}. A mention of these concepts is made by Dwyer (1967, p. 608). Thus we see that a <u>matrix element</u> specifies its position in the matrix in addition to its scalar value. In the matrix display the values of i and j are clear and the value x_{ij} is interpreted as the <u>matrix element</u> which has scalar value of x_{ij} as a scalar element. From the very definitions of the vectors X_r, X_c, X_R and X_C it is clear that each of their components has a specific position in a matrix which is a transform of the position in X.

3.2 A Matrix Derivative Transformation and its Properties

3.2.1 We begin by defining the differentiability of matrix functions with respect to a matrix variable. This generalizes the differentiability of a vector function of a vector variable; see for example Fleming (1965).

For differentiation purposes, it is an important fact that (i,j,x_{ij}) is always independent of $(i',j',x_{i',j'})$ except when $i = i'$, $j = j'$. As regards matrix elements versus scalar elements, we point out that x_{ij} and $\langle x_{ij} \rangle$ are not completely independent since x_{ij} is one of the three components of $\langle x_{ij} \rangle$. We assume that the differential operator applied to $\langle x_{ij} \rangle$ works only on x_{ij} and not on i, j. Thus

$$d\langle x_{ij} \rangle = (i,j,dx_{ij})$$

and hence the scalar value of $d\langle x_{ij} \rangle$ is the scalar dx_{ij}.

<u>Definition 3.2.1.1</u> (Differentiability). A matrix function $Y \in \mathscr{M}_{m,n}$ is differentiable at $X \in \mathscr{N}_{p,q}$ if there exists a linear transformation $L(X; H) \in \mathscr{M}_{m,n}$ such that

$$\lim_{H \to 0} \frac{Y(X+H) - Y(X) - L(X; H)}{||H||} = 0{:}m\times n \qquad (3.2.1.1)$$

where $H \in \mathscr{N}_{p,q}$, $||H|| = \text{tr } HH^T$ and L is linear in the second matrix H.

Expression (3.2.1.1) is equivalent to:

$$\lim_{H \to 0} \frac{Y_r(X+H) - Y_r(X) - L_r(X; H)}{||H||} = 0{:}mn\times 1 \qquad (3.2.1.2)$$

where Y_r and L_r are column vector representations of Y and L respectively, or

$$\lim_{H \to 0} \frac{\langle y_{ij}(X+H)\rangle - \langle y_{ij}(X)\rangle - \langle \ell_{ij}(X;H)\rangle}{\sqrt{\sum_{\alpha=1}^{p} \sum_{\beta=1}^{q} h_{\alpha\beta}^2}} = 0{:}1\times 1 \qquad (3.2.1.3)$$

for all $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; where $\langle y_{ij}(X+H)\rangle$ is the (i,j)-th matrix element of Y i.e., the element of Y which specifies the position (i,j) in the matrix in addition to the scalar value of the element.

Combining expressions (3.2.1.1) and (3.2.1.3), we get the connection between the differentiability of a matrix function Y with the differentiability of each scalar function $y_{ij}(X)$ of a matrix variable X, as defined below:

A matrix function Y is differentiable at X if and only if each of its components y_{ij}, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; is differentiable at X.

The linear matrix function L in (3.2.1.1) is called the differential of Y at X and is denoted by dY(X). This is called a 'comatrix', being an extension of 'covector' considered by Fleming (1965). If Y is differentiable at X, then the transposed matrix of the linear vector function L_r can be identified to be the matrix of the first order partial derivatives.

For further description we shall use expression (3.2.1.2).

3.2.2 Representation of a Comatrix

Corresponding to a matrix space $\mathcal{N}_{p,q}$ with an ordered basis $\mathcal{B}$, we can have a vector space W_{pq} with an ordered basis $\mathcal{B}_r$, where $\mathcal{B}_r$ is a set of column vector representations of basis matrices in the ordered basis $\mathcal{B}$.

For any linear transformation L, we denote its row-wise column vector representation by L_r.

Lemma 3.2.2.1 A matrix function $L: \mathcal{N}_{p,q} \longrightarrow \mathcal{M}_{m,n}$ defined by $L(X) = [\ell_{ij}(X)]$ is a linear transformation if and only if $L_r: W_{pq} \longrightarrow V_{mn}$ defined by $L_r(X) = [\ell_{ij}(X)]_r$ is a linear transformation for every $X \in \mathcal{N}_{p,q}$.

An application of Section 3.4 in Hoffman & Kunze (1971) yields the following theorem, which uses the concept of Lemma 3.2.2.1 above.

Theorem 3.2.2.1 Let W_{pq} and V_{mn} be two vector spaces, together with their ordered bases $\mathscr{B}_r$ and $\mathscr{B}'_r$, respectively. Then there is a one to one correspondence between the set of all linear transformations L_r defined in the Lemma 3.2.2.1 and the set of all mn×pq matrices of L_r relative to $\mathscr{B}_r$, $\mathscr{B}'_r$.

Proof: It is obvious that the set of all linear transformations L_r forms a vector space $\mathcal{U}$ of dimension pqmn. Now each of the pqmn basis matrices of $\mathcal{U}$ may be arranged as pq×mn matrices. Then the transpose of these matrices forms the set of all mn×pq matrices of L_r relative to $\mathscr{B}_r$, $\mathscr{B}'_r$. A detailed proof of this theorem like that of Theorem 11 in Hoffman & Kunze (1971, p. 87) may be given.

The above theorem is used for the identification of matrices of first order partial derivatives.

3.2.3 The Matrix of Partial Derivatives

Definition 3.2.3.1 (Partial Derivatives). If Y:m×n is a differentiable matrix function of the matrix variable X:p×q, then the (k,ℓ)-th partial derivative of $\langle y_{ij}\rangle$ with respect to the matrix elements of X is defined as

$$\lim_{h_{k\ell} \to 0} \frac{\langle y_{ij}(x_{k\ell}+h_{k\ell})\rangle - \langle y_{ij}(x_{k\ell})\rangle}{h_{k\ell}}$$

if the limit exsists, and it is denoted by $\dfrac{\partial\langle y_{ij}(X)\rangle}{\partial\langle x_{k\ell}\rangle}$, where

$i = 1, 2,\ldots, m$; $j = 1, 2,\ldots, n$; $k = 1, 2, \ldots, p$, and $\ell = 1, 2,\ldots,q$.

If, in particular, $Y = X$, then from the above definition, we obtain $\frac{\partial \langle x_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$ which is 0 when $k \neq i$ and/or $j \neq \ell$, and is 1 when $i = k$ and $j = \ell$. Here we notice that $\frac{\partial \langle x_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$ is the (i,j)-th element of $\frac{\partial X}{\partial \langle x_{k\ell} \rangle}$, and the (i,j)-th element of $\frac{\partial X^T}{\partial \langle x_{k\ell} \rangle}$ is the (i,j)-th element of $(\frac{\partial X}{\partial \langle x_{k\ell} \rangle})^T$ as discussed in Dwyer (1967). When Y is a differentiable function of X, we use chain rules to reduce $\frac{\partial Y}{\partial \langle x_{k\ell} \rangle}$ to $\sum_\alpha A_\alpha \frac{\partial X}{\partial \langle x_{k\ell} \rangle} B_\alpha + \sum_\beta C_\beta \frac{\partial X^T}{\partial \langle x_{k\ell} \rangle} D_\beta$. Thus when $Y = X^T X$, $\frac{\partial Y}{\partial \langle x_{k\ell} \rangle} = X^T \frac{\partial X}{\partial \langle x_{k\ell} \rangle} + \frac{\partial X^T}{\partial \langle x_{k\ell} \rangle} X$.

Definition 3.2.3.2 (The Matrix of Partial Derivatives). For a differentiable matrix function $Y \in \mathcal{M}_{m,n}$ of the matrix variable $X \in \mathcal{N}_{p,q}$, the matrix denoted by $\frac{\partial Y_r}{\partial X_r}$ and defined by the equation

$$\frac{\partial Y_r}{\partial X_r} = \begin{bmatrix} \frac{\partial <y_{11}>}{\partial <x_{11}>} & \cdots & \frac{\partial <y_{1n}>}{\partial <x_{11}>} & \cdots & \cdots & \frac{\partial <y_{m1}>}{\partial <x_{11}>} & \cdots & \frac{\partial <y_{mn}>}{\partial <x_{11}>} \\ \vdots & & & & & & & \\ \frac{\partial <y_{11}>}{\partial <x_{1q}>} & \cdots & \frac{\partial <y_{1n}>}{\partial <x_{1q}>} & \cdots & \cdots & \frac{\partial <y_{m1}>}{\partial <x_{1q}>} & \cdots & \frac{\partial <y_{mn}>}{\partial <x_{1q}>} \\ \vdots & & & & & & & \\ \vdots & & & & & & & \\ \frac{\partial <y_{11}>}{\partial <x_{p1}>} & \cdots & \frac{\partial <y_{1n}>}{\partial <x_{p1}>} & \cdots & \cdots & \frac{\partial <y_{m1}>}{\partial <x_{p1}>} & \cdots & \frac{\partial <y_{mn}>}{\partial <x_{p1}>} \\ \vdots & & & & & & & \\ \frac{\partial <y_{11}>}{\partial <x_{pq}>} & \cdots & \frac{\partial <y_{1n}>}{\partial <x_{pq}>} & \cdots & \cdots & \frac{\partial <y_{m1}>}{\partial <x_{pq}>} & \cdots & \frac{\partial <y_{mn}>}{\partial <x_{pq}>} \end{bmatrix} \qquad (3.2.3.1)$$

is the "Matrix of Partial Derivatives" or simply the "Matrix Derivative" in mnpq-dimensional space of all pq×mn matrices when the elements of Y are transformed to form Y_r and the elements of X to form X_r. It should be pointed out that this formulation does not give $\frac{\partial Y}{\partial <X>}$ nor $\frac{\partial <Y>}{\partial X}$. However all the partial derivatives of these expressions appear in $\frac{\partial Y_r}{\partial X_r}$, and they are identifiable, so the approach is adequate.

Similarly, the matrix derivatives $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$, and $\frac{\partial Y_c}{\partial X_c}$ are defined as matrices of partial derivatives $\frac{\partial <y_{ij}>}{\partial <x_{k\ell}>}$, uniquely ordered such that the elements of Y_r or Y_c appear as a row and those of X_r or X_c as a column. The matrix $\frac{\partial Y_c}{\partial X_c}$ was introduced by Neudecker (1969b). He denoted it by $\frac{\partial \text{ vec } Y}{\partial \text{ vec } X}$. In a paper in the same year, Tracy & Dwyer suggested all of the above four matrix derivatives. However, none of these authors provide any algebraic treatment of these matrix derivatives. Our aim here is to characterize these matrices.

The unique $(k,\ell;\ i,j)$-th element of the matrix derivatives $\frac{\partial Y_r}{\partial X_r}$, $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$, $\frac{\partial Y_c}{\partial X_c}$ is indicated by $\frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$.

Using expression (3.2.1.2), Lemma 3.2.2.1 and Theorem 3.2.2.1, we now establish a theorem which characterizes the matrix of first order partial derivatives $\frac{\partial Y_r}{\partial X_r}$.

Theorem 3.2.3.2 If $Y \in \mathscr{M}_{m,n}$ has a differential at a point $X \in \mathscr{N}_{p,q}$, then the matrix derivative $\frac{\partial Y_r}{\partial X_r}$ exists, it is unique and it is identifiable as the transpose of the matrix of the linear transformation L_r of X.

Proof: It is easy to see the existence and the uniqueness of $\frac{\partial Y_r}{\partial X_r}$. We establish the identifiability of $\frac{\partial Y_r}{\partial X_r}$ as follows:

If Y has a differential at a point X, then each of its elements $\langle y_{ij} \rangle$, $i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n$, is a differentiable scalar function of the elements $\langle x_{k\ell} \rangle$, $k = 1, 2, \ldots, p;\ \ell = 1, 2, \ldots, q$, of X, and hence

$$\ell_{ij}(dX) = \sum_{k=1}^{p} \sum_{\ell=1}^{q} \frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle} d\langle x_{k\ell} \rangle$$

$$= \left(\frac{\partial \langle y_{ij} \rangle}{\partial X_r}\right)^T dX_r \ . \qquad (3.2.3.2)$$

Collecting these components of L in the form of a column vector, we obtain

$$L_r(dX) = \begin{bmatrix} \ell_{11}(dX) \\ \vdots \\ \ell_{1n}(dX) \\ \vdots \\ \vdots \\ \ell_{m1}(dX) \\ \vdots \\ \ell_{mn}(dX) \end{bmatrix}$$

$$= \left(\frac{\partial Y_r}{\partial X_r}\right)^T dX_r \quad . \qquad (3.2.3.3)$$

Hence the matrix of the linear transformation L_r is the unique matrix formed by the mnpq elements $\frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$. This is the transpose of the first order partial matrix derivative $\frac{\partial Y_r}{\partial X_r}$.

We call L_r the <u>Matrix Derivative Transformation</u>. In other words, this is the differential of Y at X.

Some properties of L_r are as follows:

(i) L_r is linear

(ii) L_r is unique

(iii) for p = m, q = n, the absolute value of

$$\text{Det}(L_r) = \left|\frac{\partial Y_r}{\partial X_r}\right|$$

is the jacobian of the matrix transformation $X \rightarrow Y$.

If, in particular, $Y = y(X)$ is a scalar function of $X{:}p\times q$, then the matrix derivative is given by

$$\frac{\partial y}{\partial X} = \left[\frac{\partial y}{\partial x_{k\ell}}\right] :p\times q \qquad (3.2.3.4)$$

(see Dwyer (1967, p. 609)) where a typical element of $\frac{\partial y}{\partial X}$ is given by $\frac{\partial y}{\partial \langle x_{k\ell}\rangle}$. Using (3.2.1.3) and (3.2.3.2), $L = \ell_{ij} = \ell$ is a linear transformation for every $X \in \mathcal{N}_{p,q}$, and

$$\ell(dX) = \left(\frac{\partial y}{\partial X_r}\right)^T dX_r , \qquad (3.2.3.5)$$

which gives the matrix $\left(\frac{\partial y}{\partial X_r}\right)^T :pq\times 1$ of the linear transformation ℓ. The matrix derivative $\frac{\partial y}{\partial X}$ is obtained by rearranging the elements of $\frac{\partial y}{\partial X_r}$ in the matrix form and it can be identified from the following expression:

$$\begin{aligned} \ell(dX) &= \left(\frac{\partial y}{\partial X_r}\right)^T dX_r \\ &= \operatorname{tr}\left(\frac{\partial y}{\partial X}\right)^T dX , \qquad (3.2.3.6) \end{aligned}$$

using (3.2.3.4) and (3.2.3.5). Since $\frac{\partial y}{\partial X}$ is needed in many statistical applications, we are concerned with (3.2.3.6) only as long as the first order matrix derivative of a scalar function is required. Obviously ℓ is the differential dy at X.

The matrix derivatives $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$ and $\frac{\partial Y_c}{\partial X_c}$ may be similarly characterized.

The main reason for not using $\frac{\partial y}{\partial <x_{k\ell}>}$ and more general $\frac{\partial Y}{\partial <x_{k\ell}>}$ in the above discussion is that these are inadequate for characterizing simultaneously all the partial derivatives. However, the transform of the collection of these partial derivatives to vector form $\frac{\partial y}{\partial X_r}$ and to matrix form $\frac{\partial Y_r}{\partial X_r}$ respectively, are characterized as the transposes of the matrices of certain linear transformations.

3.3 A Basis Representation Theorem

In the previous section we showed that for any differentiable matrix function Y of a matrix variable X, we may observe $\left(\frac{\partial Y_r}{\partial X_r}\right)^T$ as a matrix of a certain linear transformation. For various choices of bases for the matrix space $\mathscr{L}_{pq,mn}$, we may have various representations of $\frac{\partial Y_r}{\partial X_r}$. In the present section we consider a standard basis and an ordinary basis for $\mathscr{L}_{pq,mn}$ and then show that the representations of $\frac{\partial Y_r}{\partial X_r}$ with respect to these bases have certain interrelationships.

Let

$$\mathscr{E} = \{E_{11},\ldots, E_{1,mn},\ldots \quad \ldots, E_{pq,1},\ldots, E_{pq,mn}\}$$

and

$$\mathscr{B} = \{B_{11},\ldots, B_{1,mn},\ldots \quad \ldots, B_{pq,1},\ldots, B_{pq,mn}\}$$

be the standard basis and an ordinary basis, respectively for the space $\mathscr{L}_{pq,mn}$ of all matrix derivatives $\frac{\partial Y_r}{\partial X_r}$, where $Y \in \mathscr{M}_{m,n}$ and $X \in \mathscr{N}_{p,q}$. Here we note that any $B_{\gamma\delta}$ which is an element of the arbitrary basis may be expressed uniquely in terms of the standard basis matrices $E_{\alpha\beta}$, $\alpha = 1, 2,\ldots, pq$; $\beta = 1, 2,\ldots, mn$. On the otherhand, for any $E_{\alpha\beta} \in \mathscr{E}$, there exist unique scalars $p_{\alpha\beta}^{\gamma\delta}$ (see Hoffman & Kunze (1971, p. 46)) such that

$$E_{\alpha\beta} = \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} B_{\gamma\delta} \tag{3.3.1}$$

Here we note that $E_{\alpha\beta} \in \mathscr{E}$ is the matrix with all elements equal to zero except the (α,β)-th element which is equal to 1. The subscripts α and β are defined as

$$\alpha = q(k-1)+\ell \tag{3.3.2}$$

and

$$\beta = n(i-1)+j \quad . \tag{3.3.3}$$

These are useful in specifying any desired typical element of the matrix derivative $\frac{\partial Y_r}{\partial X_r}$. It is easy to see that

$$\alpha = 1, 2,\ldots, pq \text{ for } k = 1, 2,\ldots, p;\ \ell = 1, 2,\ldots, q$$

and similarly

$$\beta = 1, 2,\ldots, mn \text{ for } i = 1, 2,\ldots, m;\ j = 1, 2,\ldots, n.$$

Obviously the matrix $\frac{\partial Y_r}{\partial X_r}$ is such that

$$\left[\frac{\partial Y_r}{\partial X_r}\right]_{\mathscr{E}} = \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle} E_{\alpha\beta}$$

$$= \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle} \left(\sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} B_{\gamma\delta}\right), \text{ using (3.3.1)},$$

$$= \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} \left(\sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} \frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle}\right) B_{\gamma\delta} \qquad (3.3.4)$$

$$= \left[\frac{\partial Y_r}{\partial X_r}\right]_{\mathscr{B}} . \qquad (3.3.5)$$

Let $z_{11}, \ldots, z_{1,mn}, \ldots \quad \ldots, z_{pq,1}, \ldots, z_{pq,mn}$ be the coordinates of the above matrix $\frac{\partial Y_r}{\partial X_r}$ in the ordered basis $\mathscr{B}$. Since these coordinates are uniquely determined, we have from (3.3.4),

$$z_{\gamma\delta} = \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} \frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle} \quad , \quad \begin{array}{l}\gamma = 1, 2, \ldots, pq \\ \delta = 1, 2, \ldots, mn\end{array} , \qquad (3.3.6)$$

where γ and δ correspond to α and β respectively, that is $\gamma = q(k-1)+\ell$ and $\delta = n(i-1)+j$.

We illustrate expression (3.3.6) by the following example.

<u>Example 3.3.1</u> Let $i = 1, 2$; $j = 1, 2$; $k = 1, 2$; $\ell = 1, 2$. Then we have $p = q = 2$; $m = n = 2$. From (3.3.2),

$\alpha = 2(1-1)+1 = 1$ when $k = 1$, $\ell = 1$

$\alpha = 2(1-1)+2 = 2$ when $k = 1$ $\ell = 2$

$\alpha = 2(2-1)+1 = 3$ when $k = 2$, $\ell = 1$

$\alpha = 2(2-1)+2 = 4$ when $k = 2$, $\ell = 2$.

Similarly, from (3.3.3),

$$\begin{aligned}\beta &= 1 \text{ when } i = 1, j = 1\\ &= 2 \text{ when } i = 1, j = 2\\ &= 3 \text{ when } i = 2, j = 1\\ &= 4 \text{ when } i = 2, j = 2.\end{aligned}$$

Hence we see that given p, q, we can determine k and ℓ uniquely for any α or for any γ . Similarly, for given m and n and for any β or δ, the proper (i,j) can be found. Thus

$$\begin{aligned} z_{11} = {} & p_{11}^{11} \frac{\partial\langle y_{11}\rangle}{\partial\langle x_{11}\rangle} + p_{12}^{11} \frac{\partial\langle y_{12}\rangle}{\partial\langle x_{11}\rangle} + p_{13}^{11} \frac{\partial\langle y_{21}\rangle}{\partial\langle x_{11}\rangle} + p_{14}^{11} \frac{\partial\langle y_{22}\rangle}{\partial\langle x_{11}\rangle} \\ & + p_{21}^{11} \frac{\partial\langle y_{11}\rangle}{\partial\langle x_{12}\rangle} + p_{22}^{11} \frac{\partial\langle y_{12}\rangle}{\partial\langle x_{12}\rangle} + p_{23}^{11} \frac{\partial\langle y_{21}\rangle}{\partial\langle x_{12}\rangle} + p_{24}^{11} \frac{\partial\langle y_{22}\rangle}{\partial\langle x_{12}\rangle} \\ & + p_{31}^{11} \frac{\partial\langle y_{11}\rangle}{\partial\langle x_{21}\rangle} + p_{32}^{11} \frac{\partial\langle y_{12}\rangle}{\partial\langle x_{21}\rangle} + p_{33}^{11} \frac{\partial\langle y_{21}\rangle}{\partial\langle x_{21}\rangle} + p_{34}^{11} \frac{\partial\langle y_{22}\rangle}{\partial\langle x_{21}\rangle} \\ & + p_{41}^{11} \frac{\partial\langle y_{11}\rangle}{\partial\langle x_{22}\rangle} + p_{42}^{11} \frac{\partial\langle y_{12}\rangle}{\partial\langle x_{22}\rangle} + p_{43}^{11} \frac{\partial\langle y_{21}\rangle}{\partial\langle x_{22}\rangle} + p_{44}^{11} \frac{\partial\langle y_{22}\rangle}{\partial\langle x_{22}\rangle} \quad , \end{aligned}$$

$$\begin{aligned} z_{12} = {} & p_{11}^{12} \frac{\partial\langle y_{11}\rangle}{\partial\langle x_{11}\rangle} + \cdots + p_{14}^{12} \frac{\partial\langle y_{22}\rangle}{\partial\langle x_{11}\rangle} + \cdots \quad \cdots + \\ & p_{41}^{12} \frac{\partial\langle y_{11}\rangle}{\partial\langle x_{22}\rangle} + \cdots + p_{44}^{12} \frac{\partial\langle y_{22}\rangle}{\partial\langle x_{22}\rangle} \ . \end{aligned}$$

Similarly z_{13}, z_{14}, ..., z_{41}, ..., z_{44} may be obtained.

These $z_{\gamma\delta}$ may be arranged as

$$\begin{bmatrix} z_{11} \\ \vdots \\ z_{14} \\ \vdots \\ \vdots \\ z_{41} \\ \vdots \\ z_{44} \end{bmatrix} = \begin{bmatrix} p^{11}_{11} & \cdots & p^{11}_{14} & \cdots & \cdots & p^{11}_{41} & \cdots & p^{11}_{44} \\ \vdots & & \vdots & & & \vdots & & \vdots \\ p^{14}_{11} & \cdots & p^{14}_{14} & \cdots & \cdots & p^{14}_{41} & \cdots & p^{14}_{44} \\ \vdots & & \vdots & & & \vdots & & \vdots \\ \vdots & & \vdots & & & \vdots & & \vdots \\ p^{41}_{11} & \cdots & p^{41}_{14} & \cdots & \cdots & p^{41}_{41} & \cdots & p^{41}_{44} \\ \vdots & & \vdots & & & \vdots & & \vdots \\ p^{44}_{11} & & p^{44}_{14} & & & p^{44}_{41} & & p^{44}_{44} \end{bmatrix} \begin{bmatrix} \frac{\partial \langle y_{11} \rangle}{\partial \langle x_{11} \rangle} \\ \vdots \\ \frac{\partial \langle y_{22} \rangle}{\partial \langle x_{11} \rangle} \\ \vdots \\ \vdots \\ \frac{\partial \langle y_{11} \rangle}{\partial \langle x_{22} \rangle} \\ \vdots \\ \frac{\partial \langle y_{22} \rangle}{\partial \langle x_{22} \rangle} \end{bmatrix} . \qquad (3.3.7)$$

Expression (3.3.7) may be put in matrix form as

$$Z_r = P\left[\frac{\partial Y_r}{\partial X_r}\right]_r \quad , \qquad (3.3.8)$$

where

$$z_{\gamma\delta} = P^{\gamma\delta}\left[\frac{\partial Y_r}{\partial X_r}\right]_r \quad . \qquad (3.3.9)$$

In general, every $Z = (z_{\gamma\delta})$ and $\frac{\partial Y_r}{\partial X_r}$ are pq×mn matrices of coordinates of $\frac{\partial Y_r}{\partial X_r}$ in the ordered bases $\mathscr{B}$ and $\mathscr{E}$ respectively, and P is the pqmn×pqmn matrix whose typical element is the scalar $p^{\gamma\delta}_{\alpha\beta}$ whenever (3.3.8) holds.

From equation (3.3.9), for $\gamma = 3$, we get $k = 2$, $\ell = 1$ and for $\delta = 2$ we get $i = 1$, $j = 2$.

Hence

$$z_{32} = \left[\frac{\partial\langle y_{12}\rangle}{\partial\langle x_{21}\rangle}\right]_{\mathscr{B}} = \sum_{\alpha=1}^{4} \sum_{\beta=1}^{4} p_{\alpha\beta}^{32} \frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle} \quad .$$

The following lemma characterizes expression (3.3.8):

<u>Lemma 3.3.1</u> Let $\{E_{\alpha\beta}\}$ and $\{B_{\alpha\beta}\}$ for $\alpha = 1, 2, \ldots, pq$; $\beta = 1, 2, \ldots, mn$, be bases of $\mathscr{L}_{pq,mn}$, defined previously.

Then the mapping

$$Z_r : U^{pqmn} \longrightarrow U^{pqmn}$$

defined by

$$Z_r\left(\left[\frac{\partial Y_r}{\partial X_r}\right]_r\right) = P\left[\frac{\partial Y_r}{\partial X_r}\right]_r \quad , \qquad (3.3.10)$$

where P is a fixed pqmn×pqmn matrix over the field of real numbers, is a well-defined function.

<u>Proof</u>: Let $\frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle}^{(1)}$ and $\frac{\partial\langle y_{ij}\rangle}{\partial\langle x_{k\ell}\rangle}^{(2)}$ be the (γ,δ)-th elements of the well-defined matrix derivatives $\frac{\partial Y_r}{\partial X_r}^{(1)}$ and $\frac{\partial Y_r}{\partial X_r}^{(2)}$.

Let

$$Z_r^{(1)} = P\left[\frac{\partial Y_r}{\partial X_r}^{(1)}\right]_r \quad \text{and} \quad Z_r^{(2)} = P\left[\frac{\partial Y_r}{\partial X_r}^{(2)}\right]_r \quad , \qquad (3.3.11)$$

and suppose further that

$$\left[\frac{\partial Y_r}{\partial X_r}^{(1)}\right]_r = \left[\frac{\partial Y_r}{\partial X_r}^{(2)}\right]_r . \qquad (3.3.12)$$

Then, since $\frac{\partial Y_r}{\partial X_r}^{(1)}$ and $\frac{\partial Y_r}{\partial X_r}^{(2)}$ are well-defined matrices, we obtain from (3.3.12)

$$\frac{\partial Y_r}{\partial X_r}^{(1)} = \frac{\partial Y_r}{\partial X_r}^{(2)} , \qquad (3.3.13)$$

which implies

$$\frac{\partial <y_{ij}>}{\partial <x_{k\ell}>}^{(1)} = \frac{\partial <y_{ij}>}{\partial <x_{k\ell}>}^{(2)} \quad \text{for} \quad \begin{array}{l} i = 1, 2, \ldots, m; \\ j = 1, 2, \ldots, n; \\ k = 1, 2, \ldots, p; \\ \ell = 1, 2, \ldots, q. \end{array} \qquad (3.3.14)$$

From (3.3.6),

$$\begin{aligned} z_{\gamma\delta}^{(1)} &= \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} \frac{\partial <y_{ij}>}{\partial <x_{k\ell}>}^{(1)} \\ &= \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} \frac{\partial <y_{ij}>}{\partial <x_{k\ell}>}^{(2)} , \text{ using (3.3.14)}, \\ &= z_{\gamma\delta}^{(2)} \quad \text{for all} \quad \begin{array}{l} \gamma = 1, 2, \ldots, pq \\ \delta = 1, 2, \ldots, mn. \end{array} \end{aligned} \qquad (3.3.15)$$

Hence from (3.3.11) and (3.3.15)

$$Z_r^{(1)} = Z_r^{(2)} . \qquad (3.3.16)$$

Hence the mapping defined in (3.3.10) is well-defined.

Now we are in a position to justify the restriction of our discussion of $\frac{\partial Y_r}{\partial X_r}$ with respect to the standard basis $\mathscr{E}$ of $\mathscr{L}_{pq,mn}$. This is clear from the following main result of this section.

<u>Theorem 3.3.1</u> Suppose that the matrix derivatives $\frac{\partial Y_r}{\partial X_r}$ with respect to the bases $\mathscr{E}$ and $\mathscr{B}$ exist. Then there exist unique scalars $p_{\alpha\beta}^{\gamma\delta}$ such that

$$\left[\frac{\partial Y_r}{\partial X_r}\right]_{\mathscr{B}} = \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} \left\{ \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \mathrm{tr}\left(\frac{\partial Y_r}{\partial X_r} E_{\beta\alpha}\right) p_{\alpha\beta}^{\gamma\delta} \right\} B_{\gamma\delta} \quad (3.3.17)$$

$$= \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} P^{\gamma\delta} \left[\frac{\partial Y_r}{\partial X_r}\right]_r B_{\gamma\delta} \quad , \quad (3.3.18)$$

where $P^{\gamma\delta}$ is the (γ,δ)-th row of the non-singular $pqmn \times pqmn$ matrix $P = (p_{\alpha\beta}^{\gamma\delta})$, $\alpha,\gamma = 1, 2,\ldots, pq$; $\beta,\delta = 1, 2,\ldots, mn$.

<u>Proof</u>: Let $\mathscr{B}$ and $\mathscr{E}$ be two ordered bases for $\mathscr{L}_{pq,mn}$. Then

$$\left[\frac{\partial Y_r}{\partial X_r}\right]_{\mathscr{B}} = \left[\frac{\partial Y_r}{\partial X_r}\right]_{\mathscr{E}}$$

$$= \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle} E_{\alpha\beta} \quad , \quad (3.3.19)$$

where k,ℓ are obtained from α and i,j are found with the help of β. Hence from (3.3.1) and (3.3.19), we obtain

$$\left[\frac{\partial Y_r}{\partial X_r}\right]_{\mathscr{B}} = \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle} \left(\sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} p_{\alpha\beta}^{\gamma\delta} B_{\gamma\delta} \right)$$

$$= \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} \left\{ \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle} p_{\alpha\beta}^{\gamma\delta} \right\} B_{\gamma\delta} \quad (3.3.20)$$

$$= \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} \left\{ \sum_{\alpha=1}^{pq} \sum_{\beta=1}^{mn} \mathrm{tr}\left(\frac{\partial Y_r}{\partial X_r} E_{\beta\alpha}\right) p_{\alpha\beta}^{\gamma\delta} \right\} B_{\gamma\delta} \quad ,$$

where $tr\left(\frac{\partial Y_r}{\partial X_r} E_{\beta\alpha}\right)$ is the α-th element of the β-th column of $\frac{\partial Y_r}{\partial X_r}$. This proves (3.3.17)

Using (3.3.6), (3.3.9) and (3.3.20), we have

$$\left[\frac{\partial Y_r}{\partial X_r}\right] = \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} z_{\gamma\delta} B_{\gamma\delta}$$

$$= \sum_{\gamma=1}^{pq} \sum_{\delta=1}^{mn} p^{\gamma\delta} \left[\frac{\partial Y_r}{\partial X_r}\right]_r B_{\gamma\delta} \quad ,$$

which proves (3.3.18).

In particular, for a scalar function y = f(X), Wrobleski (1963, Theorem 2.2.2) has discussed a certain basis representation of $\frac{\partial f}{\partial X}$. Our approach is slightly different and features more general cases.

With the help of Theorem 3.3.1, we can obtain formulae for $\left[\frac{\partial Y_r}{\partial X_c}\right]_{\mathscr{B}}$, $\left[\frac{\partial Y_c}{\partial X_r}\right]_{\mathscr{B}}$ and $\left[\frac{\partial Y_c}{\partial X_c}\right]_{\mathscr{B}}$ using various interrelationships among matrix derivatives $\frac{\partial Y_r}{\partial X_r}$, $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$ and $\frac{\partial Y_c}{\partial X_c}$. These interrelationships are considered in Section 3.7.

Thus, from Theorem 3.3.1, it follows that for any of the above matrix derivatives, we may very conveniently choose its representation with respect to the standard basis $\mathscr{E}$. This is possible, since the representation of these with respect to any arbitrary basis $\mathscr{B}$ may be described by the corresponding matrix derivatives and the matrices $\{E_{\alpha\beta}\}$ and $\{B_{\gamma\delta}\}$ of the bases $\mathscr{E}$ and $\mathscr{B}$ respectively.

However, it may be desirable to point out the trivial modifications needed in obtaining the (α,β)-th elements of $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$ and $\frac{\partial Y_c}{\partial X_c}$ respectively. These are given below:

(i) if $\frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$ is the (α,β)-th element of $\frac{\partial Y_r}{\partial X_c}$, then

$\alpha = k+p(\ell-1)$ and $\beta = n(i-1)+j$;

(ii) if $\frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$ is the (α,β)-th element of $\frac{\partial Y_c}{\partial X_r}$, then

$\alpha = q(k-1)+\ell$ and $\beta = i+m(j-1)$;

(iii) if $\frac{\partial \langle y_{ij} \rangle}{\partial \langle x_{k\ell} \rangle}$ is the (α,β)-th element of $\frac{\partial Y_c}{\partial X_c}$, then

$\alpha = k+p(\ell-1)$ and $\beta = i+m(j-1)$.

In all the above three cases the knowledge of α and β is enough to determine k,ℓ and i,j uniquely if these are required.

3.4 General Procedure for Identification of Matrix Derivatives

We use the following theorem, where $dX = [dx_{ij}]$.

Theorem 3.4.1 For a matrix function $Y = F(X)$, $X{:}m\times n$

$$\text{(i)} \quad (dX^T)_r = I_{(n)} dX_r \tag{3.4.1}$$

$$\text{(ii)} \quad dY_r = P dX_r \implies \frac{\partial Y_r}{\partial X_r} = P^T . \tag{3.4.2}$$

Proof: (i) Obvious from equation (2.6.3.10).

(ii) Since $d\langle y_{k\ell}\rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial \langle y_{k\ell}\rangle}{\partial \langle x_{ij}\rangle} d\langle x_{ij}\rangle$, we have,

by ordering the partial derivatives $\frac{\partial \langle y_{k\ell}\rangle}{\partial \langle x_{ij}\rangle}$,

$$dY_r = \left(\frac{\partial Y_r}{\partial X_r}\right)^T dX_r$$

which, together with $dY_r = P dX_r$, implies $\frac{\partial Y_r}{\partial X_r} = P^T$.

The identification of $\frac{\partial Y_r}{\partial X_r}$ from $dY_r = P dX_r$ in (ii) above would be more direct if Definition 3.2.3.2 had required the elements of Y_r to appear as a column and those of X_r as a row, for then

$$dY_r = P dX_r \quad \text{would imply} \quad \frac{\partial Y_r}{\partial X_r} = P$$

in place of P^T . Although this could save transposition when matrix derivatives become complicated, we would go along with Definition 3.2.3.2 to retain consistency with Neudecker (1969b) and Tracy & Dwyer (1969).

In Section 3.11, the above concepts are generalized and then they become useful in differentiating partitioned matrix functions.

A very general matrix function, e.g.,

$$Y = A(X^{-1})^{\alpha}B(X^{T})^{\gamma}CX^{\alpha}D(X^{-T})^{\delta}E$$

of a matrix variable X can be differentiated by using Theorem 3.4.1 and the fact that

$$dX^{\alpha} = \sum_{k=0}^{\alpha-1} X^{k}(dX)X^{\alpha-k-1} \text{ , for a positive integer } \alpha$$

and $dX^{\alpha} = -\sum_{k=1}^{-\alpha} X^{-k}(dX)X^{\alpha+k-1}$, for a negative integer α.

This is achieved in the following theorem:

<u>Theorem 3.4.2</u> For any differentiable matrix function $Y = F(X)$ of a matrix variable $X{:}m{\times}n$ involving ordinary matrix products, and for matrices A_i, B_i, C_j and D_j $(i = 1, 2,\ldots, s;$ $j = 1, 2,\ldots, t)$ which are matrix functions of X of appropriate orders, we have

$$dY = \sum_{1}^{s} A_i(dX)B_i + \sum_{1}^{t} C_j(dX^{T})D_j \tag{3.4.3}$$

and

$$\frac{\partial Y_r}{\partial X_r} = \sum_{1}^{s} (A_i^{T} \otimes B_i) + I_{(m)} \sum_{1}^{t} (C_j^{T} \otimes D_j) \ . \tag{3.4.4}$$

<u>Proof</u>: Obviously, Y can be expressed as a sum of terms involving ordinary matrix multiplication, each term containing finite powers of X, X^{T} and X^{-1} in a conformable fashion. Then the differential dY can be expressed as the sum of $s = u+v$

terms $\sum_1^s A_i(dX)B_i$ and a sum of t terms $\sum_1^t C_j(dX^T)D_j$; where non-negative integers u, v and t are the aggregates of the powers of X, X^{-1} and X^T, respectively, occurring in the explicit representation of a matrix function Y of X.

Applying property (A.1.6), the results of Section 2.6.3 and Theorem 3.4.1 (i), we get

$$dY_r = \left[\sum_1^s (A_i \otimes B_i{}^T) + \sum_1^t (C_j \otimes D_j{}^T) I_{(n)}\right] dX_r \quad .$$

This proves (3.4.4) on using Theorem 3.4.1(ii), (2.4.2) and (A.1.7).

Example 3.4.1 Suppose we have $X{:}m{\times}n$, and

$$Y = X^T P X X^T Q X^{-1} R X^T S \quad .$$

Then

$$dY = \sum_1^2 A_i(dX)B_i + \sum_1^3 C_j(dX^T)D_j$$

where $A_1 = X^T P$, $A_2 = -X^T P X X^T Q X^{-1}$, $B_1 = X^T Q X^{-1} R X^T S$, $B_2 = X^{-1} R X^T S$, $C_1 = I$, $C_2 = X^T P X$, $C_3 = X^T P X X^T Q X^{-1} R$, $D_1 = P X X^T Q X^{-1} R X^T S$, $D_2 = Q X^{-1} R X^T S$, $D_3 = S$, and $s = 2$, $u = 1$, $v = 1$, $t = 3$.

Further,

$$dY_r = \left[\sum_1^2 (A_i \otimes B_i{}^T) + \sum_1^3 (C_j \otimes D_j{}^T) I_{(n)}\right] dX_r$$

which implies

$$\frac{\partial Y_r}{\partial X_r} = \sum_1^2 (A_i{}^T \otimes B_i) + I_{(m)} \sum_1^3 (C_j{}^T \otimes D_j) \quad .$$

In the spirit of Theorem 3.4.2, we can form the following table of formulae for all forms of matrix derivatives.

TABLE 3.4.1 FORMULAE FOR $\frac{\partial Y_r}{\partial X_r}$, $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$, $\frac{\partial Y_c}{\partial X_c}$

WHEN $dY = \sum_i A_i(dX)B_i + \sum_j C_j(dX^T)D_j$.

Partial derivative of → / With respect to ↓	Y_r	Y_c
X_r	$\sum_i(A_i^T \otimes B_i) + I_{(m)} \sum_j(C_j^T \otimes D_j)$	$I_{(m)} \sum_i(B_i \otimes A_i^T) + \sum_j(D_j \otimes C_j^T)$
X_c	$I_{(n)} \sum_i(A_i^T \otimes B_i) + \sum_j(C_j^T \otimes D_j)$	$\sum_i(B_i \otimes A_i^T) + I_{(n)} \sum_j(D_j \otimes C_j^T)$

Table 3.4.1 is a slight variation of the table presented by Tracy & Dwyer (1969, Table 2).

Theorem 3.4.2 may be compared with Dwyer (1967, p. 612), Neudecker (1969b, p. 956) and Tracy & Dwyer (1969, p. 1580).

For differentiable scalar functions of matrices, a theorem analogous to Theorem 3.4.2 is the following:

Theorem 3.4.3 If f(X) is a differentiable scalar function of a matrix variable X, then

$$df(X) = tr[\Sigma A_i(dX)B_i + \Sigma C_j(dX^T)D_j] \tag{3.4.5}$$

and $$\frac{\partial f(X)}{\partial X} = \Sigma A_i{}^T B_i{}^T + \Sigma D_j C_j \ , \tag{3.4.6}$$

where A_i, B_i, C_j and D_j may be some matrix functions of X.

Proof: Since dtrY = trdY, (3.4.5) is obvious.

Expression (3.4.6) follows from

$$df(X) = tr(\frac{\partial f}{\partial X})^T dX \text{ , using (3.2.3.6);}$$

and $tr[\Sigma A_i(dX)B_i + \Sigma C_j(dX^T)D_j] = tr[\Sigma A_i{}^T B_i{}^T + \Sigma D_j C_j]^T(dX)$,

using properties of traces of matrices.

Theorem 3.4.3 extends a result in Neudecker (1969b). Tracy & Dwyer (1969, (4.2), p. 1581), using J-type and K-type matrices, have derived expression (3.4.6).

3.5 Illustration of the General Identification Theorem

The general procedure for identifying derivative matrices may be successfully applied to more general matrix functions, for example, functions involving any finite powers which may be positive or negative integers. It would be convenient to arrange $\frac{\partial Y_r}{\partial X_r}$ for more general matrix functions Y = F(X), X:m×n or X:m×m in a tabular form.

TABLE 3.5.1 SOME FORMULAE FOR $\frac{\partial Y_r}{\partial X_r}$.

Y	$\frac{\partial Y_r}{\partial X_r}$
CXX^TAX^TDX	$C^T \otimes X^TAX^TDX + I_{(m)}(X^TC^T \otimes AX^TDX) +$ $I_{(m)}(A^TXX^TC^T \otimes DX) + D^TXA^TXX^TC^T \otimes I_n$
$XAX^{-1}BX^T$	$I_m \otimes AX^{-1}BX^T - (X^{-T}A^TX^T \otimes X^{-1}BX^T)$ $+ I_{(m)}(B^TX^{-T}A^TX^T \otimes I_m)$
$X^TAX^{-T}BX$	$I_{(m)}(I_m \otimes AX^{-T}BX) - I_{(m)}(X^{-1}A^TX \otimes X^{-T}BX)$ $+ (B^TX^{-1}A^TX \otimes I_m)$
X^α, α positive integer	$\sum_{0}^{\alpha-1} (X^T)^i \otimes X^{\alpha-i-1}$
$(X^T)^\alpha$, α positive integer	$I_{(m)} \sum_{0}^{\alpha-1} X^i \otimes (X^T)^{\alpha-i-1}$
X^α, α negative integer	$-\sum_{1}^{-\alpha} (X^T)^{-i} \otimes X^{\alpha+i-1}$
$(X^T)^\alpha$, α negative integer	$-I_{(m)} \sum_{1}^{-\alpha} X^{-i} \otimes (X^T)^{\alpha+i-1}$
e^X, X square matrix	$\frac{1}{p!} \sum_{p=1}^{\infty} \sum_{i=0}^{p-1} (X^T)^i \otimes X^{p-i-1}$

Tracy & Dwyer (1969, Table 3) have considered other particular matrix functions and have obtained formulae for $\frac{\partial Y_r}{\partial X_r}$.

3.6 Algebraic Structure of Matrix Derivative Transformations

The following theorem describes the algebraic structure of the matrix derivative transformation $\frac{\partial Y_r}{\partial X_r}$:

Theorem 3.6.1 Let Z and W be differentiable matrix functions of an independent matrix variable X, and let A, B, C and D be constant matrices. Then

$$\frac{\partial (AZB+CWD)_r}{\partial X_r} = \frac{\partial Z_r}{\partial X_r}(A^T \otimes B) + \frac{\partial W_r}{\partial X_r}(C^T \otimes D), \tag{3.6.1}$$

$$\frac{\partial (AZBWC)_r}{\partial X_r} = \frac{\partial Z_r}{\partial X_r}(A^T \otimes BWC) + \frac{\partial W_r}{\partial X_r}(B^T Z^T A^T \otimes C). \tag{3.6.2}$$

Proof: It is obvious that the matrix functions AZB+CWD and AZBWC are differentiable at X. Further we have:

$$d(AZB+CWD) = A(dZ)B + C(dW)D$$

$$\Longrightarrow d(AZB+CWD)_r = (A \otimes B^T)dZ_r + (C \otimes D^T)dW_r$$

$$\Longrightarrow \left[\frac{\partial (AZB+CWD)_r}{\partial X_r}\right]^T dX_r = \left[(A \otimes B^T)\left(\frac{\partial Z_r}{\partial X_r}\right)^T + (C \otimes D^T)\left(\frac{\partial W_r}{\partial X_r}\right)^T\right] dX_r$$

which leads to (3.6.1), and similarly

$$\left[\frac{\partial (AZBWC)_r}{\partial X_r}\right]^T dX_r = \left[(A \otimes C^T W^T B^T)\left(\frac{\partial Z_r}{\partial X_r}\right)^T + (AZB \otimes C^T)\left(\frac{\partial W_r}{\partial X_r}\right)^T\right] dX_r$$

leads to (3.6.2).

Remark: Combining (3.6.1) and (3.6.2), we have

$$\frac{\partial (AZBWC+DUQ)_r}{\partial X_r} = \frac{\partial Z_r}{\partial X_r}(A^T \otimes BWC) + \frac{\partial W_r}{\partial X_r}(B^T Z^T A^T \otimes C) + \frac{\partial U_r}{\partial X_r}(D^T \otimes Q) \ ,$$

where U is an additional differentiable matrix function and Q is a constant matrix.

We demonstrate the use of Theorem 3.6.1 in providing a formal proof of the following obvious result.

Theorem 3.6.2 Let A be a constant matrix. Then, for a matrix variable X, we have

$$\frac{\partial A_r}{\partial X_r} = 0 \ ,$$

where 0 is a zero matrix.

Proof: For a non-singular matrix function Y of X such that AY is defined, we have

$$A = AYY^{-1} \ .$$

Taking differentials,

$$dA = A(dY)Y^{-1} - AYY^{-1}(dY)Y^{-1} \ ,$$

from which, on using Theorem 3.6.1, we get

$$\frac{\partial A_r}{\partial X_r} = \frac{\partial Y_r}{\partial X_r}\left[(A^T \otimes Y^{-1}) - (A^T \otimes Y^{-1})\right] = 0 \ .$$

3.7 Some Interrelationships Among Derivative Matrices

Tracy & Dwyer (1969, p. 1579) provide expressions for $\frac{\partial Y_r}{\partial X_r}$, $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$ and $\frac{\partial Y_c}{\partial X_c}$ separately when $\frac{\partial Y}{\partial \langle X \rangle} = AJB$ or $\frac{\partial Y}{\partial \langle X \rangle} = CJ^TD$. In order to justify the fact that the study of the calculus of only one form of the above four matrix derivatives is sufficient to supply the useful information about the remaining three forms, we present their various interrelationships diagramatically.

Let Y:m×n be any differentiable matrix function of a matrix variable X:p×q. Then all possible interrelationships are available from Figure 3.7.1, where $I_{(k)}$ before/after an arrow head indicates pre-multiplication /post-multiplication to the original matrix derivative. For example

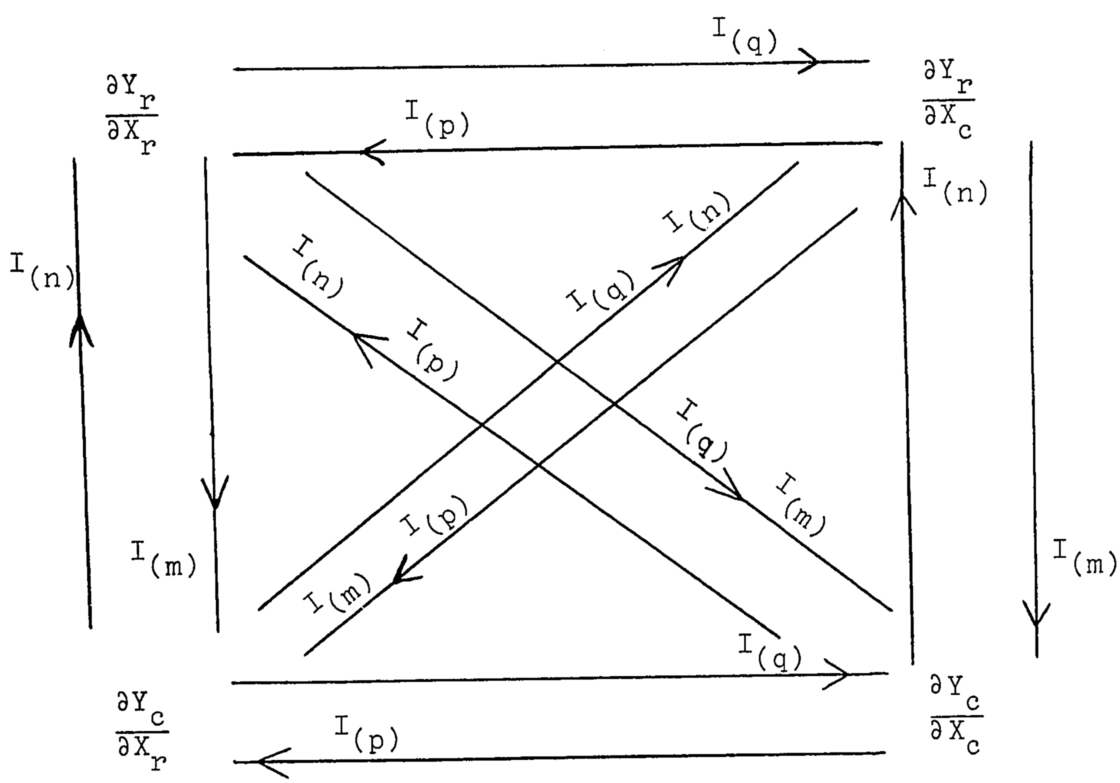
$$\frac{\partial Y_r}{\partial X_r} \xrightarrow{I_{(q)}} \frac{\partial Y_r}{\partial X_c} \Longrightarrow \frac{\partial Y_r}{\partial X_c} = I_{(q)} \frac{\partial Y_r}{\partial X_r} ,$$

$$\frac{\partial Y_c}{\partial X_c} \xleftarrow{I_{(m)}} \frac{\partial Y_r}{\partial X_c} \Longrightarrow \frac{\partial Y_c}{\partial X_c} = \frac{\partial Y_r}{\partial X_c} I_{(m)} ,$$

and

$$\frac{\partial Y_r}{\partial X_r} \xrightarrow{I_{(q)}\ I_{(m)}} \frac{\partial Y_c}{\partial X_c} \Longrightarrow \frac{\partial Y_c}{\partial X_c} = I_{(q)} \frac{\partial Y_r}{\partial X_r} I_{(m)} .$$

FIGURE 3.7.1 INTERRELATIONSHIPS AMONG MATRIX DERIVATIVES



Entries in the table presented by Tracy & Dwyer (1969, Table 2) can be obtained, in particular, by using Figure 3.7.1. For example,

$$\frac{\partial Y_r}{\partial X_r} = \frac{\partial Y_c}{\partial X_r} I_{(n)} = I_{(p)}(B\otimes A^T)I_{(n)}$$

$$= A^T\otimes B \quad \text{using Theorem 2.5.1 (iii), which}$$

may also be obtained directly by applying the general identification procedure for $\frac{\partial Y_r}{\partial X_r}$. Similarly $\frac{\partial Y_r}{\partial X_c}$ and $\frac{\partial Y_c}{\partial X_c}$ can be obtained with the help of $\frac{\partial Y_c}{\partial X_r}$ by using Figure 3.7.1.

3.8 Identification of Mixed Partial Matrix Derivatives for Scalar Functions of Matrix Variables

In this section, we establish two matrix differentiation theorems which are useful in tackling multivariate statistical problems. We need the following results which are straightforward:

$$\operatorname{tr}(XAY) = (X_r^T)(I\otimes A)(Y^T)_r \tag{3.8.1}$$

$$d^2\operatorname{tr}F(X) = \operatorname{tr}d^2F(X) \tag{3.8.2}$$

$$d^2f(X,Y) = g(dX,dY) \tag{3.8.3}$$

Definition 3.8.1 (Mixed Partial Matrix Derivative). Let f be a differentiable scalar function of matrix variables X, Y,..., Z. Then a Mixed Partial Matrix Derivative $\nabla^2_{Y_r,X_r}f$ of f(X, Y,..., Z) with respect to any pair of transformed vector variables, e.g. with respect to X_r:mn×1 and Y_r:pq×1 is a matrix $\left[\frac{\partial^2 f}{\partial \langle y_{k\ell}\rangle \partial \langle x_{ij}\rangle}\right]$ of mixed partial matrix element derivatives, uniquely ordered as below:

$$\nabla^2_{Y_r,X_r} f(X,Y) = \frac{\partial}{\partial Y_r}\left(\frac{\partial f}{\partial X_r}\right)$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial\langle y_{11}\rangle\partial\langle x_{11}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{11}\rangle\partial\langle x_{1n}\rangle}\cdots & \cdots\frac{\partial^2 f}{\partial\langle y_{11}\rangle\partial\langle x_{m1}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{11}\rangle\partial\langle x_{mn}\rangle} \\ \vdots & \\ \frac{\partial^2 f}{\partial\langle y_{1q}\rangle\partial\langle x_{11}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{1q}\rangle\partial\langle x_{1n}\rangle}\cdots & \cdots\frac{\partial^2 f}{\partial\langle y_{1q}\rangle\partial\langle x_{m1}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{1q}\rangle\partial\langle x_{mn}\rangle} \\ \vdots & \\ \vdots & \\ \frac{\partial^2 f}{\partial\langle y_{p1}\rangle\partial\langle x_{11}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{p1}\rangle\partial\langle x_{1n}\rangle}\cdots & \cdots\frac{\partial^2 f}{\partial\langle y_{p1}\rangle\partial\langle x_{m1}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{p1}\rangle\partial\langle x_{mn}\rangle} \\ \vdots & \\ \frac{\partial^2 f}{\partial\langle y_{pq}\rangle\partial\langle x_{11}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{pq}\rangle\partial\langle x_{1n}\rangle}\cdots & \cdots\frac{\partial^2 f}{\partial\langle y_{pq}\rangle\partial\langle x_{m1}\rangle}\cdots\frac{\partial^2 f}{\partial\langle y_{pq}\rangle\partial\langle x_{mn}\rangle} \end{bmatrix}$$

$$= \begin{matrix} & x_{11}\cdots x_{1n}\cdots\;\cdots x_{m1}\cdots x_{mn} \\ \begin{matrix} y_{11} \\ \vdots \\ y_{1q} \\ \vdots \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pq} \end{matrix} & \left[\begin{matrix} \\ \\ \\ \frac{\partial^2 f}{\partial\langle y_{k\ell}\rangle\partial\langle x_{ij}\rangle} \\ \\ \\ \\ \end{matrix}\right] \end{matrix} .$$

Using (3.4.1) and (3.8.1), we establish the general identification theorem which is our main result:

Theorem 3.8.1 For any differentiable scalar function $f(X,Y)$ of $X{:}m\times n$, $Y{:}p\times q$, and $I_{(k)}$ as in Definition 2.4.1,

$$d^2f = \operatorname{tr} A(dX)B(dY)C \implies \nabla^2_{Y_r,X_r} f = I_{(p)}(CA\otimes B^T). \quad (3.8.4)$$

Proof: Using expression (3.8.1), we obtain

$$\begin{aligned}\operatorname{tr} A(dX)B(dY)C &= \{A(dX)\}_r^T(I\otimes B)\{(dY)C\}_r^T \\ &= \{(A\otimes I)dX_r\}^T(I\otimes B)\{(C^T\otimes I)(dY^T)_r\} \\ &= (dX_r)^T(A^T\otimes I)(I\otimes B)(C^T\otimes I)I_{(q)}dY_r \\ &= (dX_r)^T(A^TC^T\otimes B)I_{(q)}dY_r \text{ , using (A.1.6)} \\ &= (dY_r)^T I_{(p)}(CA\otimes B^T)dX_r \text{ ,} \quad (3.8.5)\end{aligned}$$

using (A.2.1) and (2.4.2).

Also, the mixed differential of f with respect to the matrix variables $X{:}m\times n$ and $Y{:}p\times q$, respectively, is expressed as

$$d^2f = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{p}\sum_{\ell=1}^{q} \frac{\partial^2 f}{\partial\langle y_{k\ell}\rangle\partial\langle x_{ij}\rangle}\, d\langle y_{k\ell}\rangle d\langle x_{ij}\rangle$$

$$= (dY_r)^T(\nabla^2_{Y_r,X_r} f)dX_r \text{ .} \quad (3.8.6)$$

From expressions (3.8.5) and (3.8.6), we obtain

$$\nabla^2_{Y_r,X_r} f = I_{(p)}(CA\otimes B^T),$$

which proves the required result.

For a pair of matrix variables X and Y, there are seven other possible forms of Theorem 3.8.1. These we give as corollaries. (In fact, any of these eight results may be treated as a theorem and the others derived as corollaries). Let X, Y and f(X,Y) be as in Theorem 3.8.1. Then

Corollary 3.8.1

$$d^2 f = \text{tr } A(dY)B(dX)C \Longrightarrow \nabla^2_{Y_r, X_r} f = (A^T C^T \otimes B) I_{(n)}. \quad (3.8.7)$$

Corollary 3.8.2

$$d^2 f = \text{tr } A(dX^T)B(dY)C \Longrightarrow \nabla^2_{Y_r, X_r} f = B^T \otimes CA. \quad (3.8.8)$$

Corollary 3.8.3

$$d^2 f = \text{tr } A(dY)B(dX^T)C \Longrightarrow \nabla^2_{Y_r, X_r} f = A^T C^T \otimes B. \quad (3.8.9)$$

Corollary 3.8.4

$$d^2 f = \text{tr } A(dX)B(dY^T)C \Longrightarrow \nabla^2_{Y_r, X_r} f = CA \otimes B^T. \quad (3.8.10)$$

Corollary 3.8.5

$$d^2 f = \text{tr } A(dY^T)B(dX)C \Longrightarrow \nabla^2_{Y_r, X_r} f = B \otimes A^T C^T. \quad (3.8.11)$$

Corollary 3.8.6

$$d^2 f = \text{tr } A(dX^T)B(dY^T)C \Longrightarrow \nabla^2_{Y_r, X_r} f = (CA \otimes B^T) I_{(n)}. \quad (3.8.12)$$

Corollary 3.8.7

$$d^2 f = \text{tr } A(dY^T)B(dX^T)C \Longrightarrow \nabla^2_{Y_r, X_r} f = I_{(p)}(A^T C^T \otimes B). \quad (3.8.13)$$

Theorem 3.8.1 and Corollary 3.8.4 are also available in Tracy & Singh (1971a).

We now discuss some important particular cases of Theorem 3.8.1 and its corollaries. The following theorem is a generalization of a result in Fleming (1965, p. 47):

Theorem 3.8.2 If, in the neighbourhood of X and Y, all of the second order partial derivatives $\frac{\partial^2 f}{\partial \langle y_{k\ell}\rangle \partial \langle x_{ij}\rangle}$, $(i,j) \neq (k,\ell)$, exist and are continuous, then

$$\nabla^2_{Y_r,X_r} f = (\nabla^2_{X_r,Y_r} f)^T. \tag{3.8.14}$$

Proof: From the above definition, the

$$\begin{aligned}(k,\ell;\ i,j)\text{-th element of } (\nabla^2_{X_r,Y_r} f)^T &= \frac{\partial^2 f}{\partial \langle x_{ij}\rangle \partial \langle y_{k\ell}\rangle} \\ &= \frac{\partial^2 f}{\partial \langle y_{k\ell}\rangle \partial \langle x_{ij}\rangle} \\ &\quad \text{(cf. Fleming (1965, p. 47, Theorem 3))} \\ &= \text{the } (k,\ell;\ i,j)\text{-th element of } \nabla^2_{Y_r,X_r} f.\end{aligned}$$

Hence the result.

An interesting special case of the above theorem arises when Y = X. In this case $\nabla^2_{X_r,X_r} f$ is a symmetric matrix of second order partial derivatives, called the Hessian matrix of f at X.

If in expressions (3.8.4) and (3.8.7), we put $Y = X{:}m\times n$, then either $I_{(m)}(CA\otimes B^T)$ or $(A^TC^T\otimes B)I_{(n)}$ may be treated as the required Hessian matrix of f at X, since

$$(A^TC^T\otimes B)I_{(n)} = [(A^TC^T\otimes B)I_{(n)}]^T = I_{(m)}(CA\otimes B^T).$$

From the other expressions of the above corollaries, three additional Hessian matrices may similarly be obtained by specifying Y = X:m×n.

The results of Neudecker ((1967, p. 103), (1969b, p. 957)) may be obtained as particular cases of Theorem 3.8.1, Corollaries 3.8.1-3.8.7 and Theorem 3.8.2.

Let Y = X in (3.8.10). Then

$$d^2f = \operatorname{tr} A(dX)B(dX^T)C \Longrightarrow \nabla^2_{X_r,X_r} f = CA \otimes B^T$$
$$= A^TC^T \otimes B,$$
using Theorem 3.8.2.

This is equivalent to:

$$d^2f = \operatorname{tr} A(dX)B(dX^T) \Longrightarrow \left[\frac{\partial^2 f}{\partial\langle x_{ij}\rangle \partial\langle x_{k\ell}\rangle}\right] = A^T \otimes B,$$

where $A^T \otimes B$ is symmetric matrix, as given by Neudecker (1967, p. 103). In our terminology this can be written as :

$$\nabla^2_{X_r,X_r} f = \left[\frac{\partial^2 f}{\partial\langle x_{ij}\rangle \partial\langle x_{k\ell}\rangle}\right] = A^T \otimes B,$$

for the symmetric matrix $A^T \otimes B$.

We shall need the following results which are immediately obtained from Figure 3.7.1:

$$\nabla^2_{Y_c,X_c} f = I_{(q)}\left(\nabla^2_{Y_r,X_r} f\right) I_{(m)} \qquad (3.8.15)$$

$$\nabla^2_{X_c,Y_c} f = I_{(n)}\left(\nabla^2_{X_r,Y_r} f\right) I_{(p)} \; . \qquad (3.8.16)$$

Now we can derive the result of Neudecker (1969b, p.957), as a particular case of (3.8.8) or (3.8.11) by specifying Y = X and using (3.8.15) or (3.8.16).

Some applications of the results of this section are given in Chapter IV.

3.9 Testing Extrema of Matrix Functions

In this section we apply formulae for first and second order matrix derivatives to find extrema (minima and maxima) of differentiable scalar functions of matrices. We consider unconstrained and constrained extremum problems and establish a set of necessary and sufficient conditions for locating extrema for these problems.

3.9.1 Unconstrained Extrema

Theorem 3.9.1 (Tracy & Dwyer (1969)). Let $y = f(X)$ be a differentiable scalar function of a matrix variable $X{:}m{\times}n$. A necessary condition for an extremum of y at X^o is that

$$\frac{\partial y}{\partial X} = 0 \tag{3.9.1}$$

when $X = X^o$.

A sufficient condition for $f(X^o)$ to be a minimum is that

$$\nabla^2_{X_r, X_r} y \ , \tag{3.9.2}$$

when evaluated at X^o, be positive definite; and

a sufficient condition for $f(X^o)$ to be a maximum is that

$$\nabla^2_{X_r, X_r} y \ , \tag{3.9.3}$$

when evaluated at X^o, be negative definite.

Proof: The approach followed by Gillespie (1951) for a function of a vector variable may be extended to prove the above theorem.

3.9.2 Constrained Extrema

For a matrix variable $X:m\times n$, let $y = f(X)$ and $G(X) = 0:p\times q$, $pq < mn$, be differentiable matrix functions of X. Let $\Lambda:q\times p$ be a matrix of Lagrangian multipliers. We define a function

$$z(X,\Lambda) = y + \text{tr}\Lambda G . \qquad (3.9.4)$$

We then proceed to find the extrema of z at (X^o, Λ^o) which provides us with the extrema of y at X^o subject to pq constraints $G(X) = 0$.

Theorem 3.9.2 A set of necessary conditions for a constrained extremum of y at X^o is that

$$\frac{\partial z}{\partial X} = 0 \qquad (3.9.5)$$

and $G(X) = 0$ (3.9.6)

when $X = X^o$.

A sufficient condition for $f(X^o)$ to be a constrained minimum (maximum) is that

$$\nabla^2_{X_r, X_r} z \; , \qquad (3.9.7)$$

when evaluated at X^o, be positive definite (negative definite)

subject to $(dX_r)^T \left(\frac{\partial G_r}{\partial X_r}\right) = 0:1\times pq.$ (3.9.8)

Proof: Again the proof of this theorem is a generalization of that of Gillespie (1951, §20) to scalar function y of a matrix variable.

Tracy & Dwyer (1969) have considered a particular case of the above theorem applied to matrix derivatives. In their case X and Λ are vectors and hence are not applicable to general multivariate models where the argument X is a matrix.

For detailed proofs of the above particular cases, we refer to Gillespie (1951). The sufficient condition for constrained maxima for any differentiable scalar function of a vector variable given by Fleming (1965, p. 134, Ex. 11) seems to have some errors. In his notation, $Q(\underline{x}_o,\underline{h}) \leq 0$ for a maximum and $Q(\underline{x}_o,\underline{h}) < 0$ for a strict maximum, whereas he has these inequalities reversed. If $y = f(\underline{x})$ is a linear differentiable function of a vector variable, then the extremum problems become particular cases of the work of Gillespie (1951), and expressions for such problems are given by Goldberger (1964).

Expressions (3.9.1)-(3.9.8) are very general and are useful for more general multivariate extremum problems. Some of these applications are given in Chapter IV.

3.10 Partitioned Matrix Differentiation

In this section, we present an introduction to the matrix differential calculus which is based on matrix functions involving partitioned matrices. In a recent paper, Tracy & Singh (1970b) discussed the matrix differentiation of linear partitioned matrix functions with the help of the matrix product $\circledast$. Here our main attempt is to develop a matrix differential calculus of some general partitioned matrix functions which parallels that of ordinary (non-partitioned) matrix functions mentioned in previous sections of this dissertation. Some extensions of the work of Tracy & Singh (1970b, 1971b), together with its introductory details are presented in this section.

Auxiliary matrices ${}_{\{m\}}I$, $I_{\{n\}}$; the partitioned matrix product $\textcircled{\pi}$ and a procedure for obtaining Y_R and Y_C as functions of X_R and/or X_C are the basic tools in differentiating the transforms of the partitioned matrix functions $Y = F(X)$. This generalizes the basic procedure of non-partitioned matrix differentiation treated by Dwyer (1967), Neudecker (1969b) and Tracy & Dwyer (1969).

An econometric application of such differentiation is provided in Chapter IV.

A very general procedure for identifying the partial matrix derivatives for some partitioned matrix functions is given by the following theorem:

<u>Theorem 3.10.1</u> For any differentiable partitioned matrix function $Y = F(X)$, where $X{:}p\times q$ is partitioned into m row blocks and n column blocks, we have

$$\text{(i)} \qquad dY_R = P dX_R \Longrightarrow \frac{\partial Y_R}{\partial X_R} = P^T \tag{3.10.1}$$

$$\text{(ii)} \qquad (dX^T)_R = (I_{\{n\}}) dX_R = dX_C \tag{3.10.2}$$

where dX_R and dX_C are to be regarded as abbreviations of $(dX)_R$ and $(dX)_C$.

<u>Proof</u>: (i) It can be readily seen that the differential of $Y^{k\ell}$ may be arranged as

$$dY^{k\ell}{}_r = \left(\frac{\partial Y^{k\ell}{}_r}{\partial X_R}\right) dX_R \quad .$$

A unique ordering of the submatrices of partial derivatives

$\frac{\partial Y^{k\ell}{}_r}{\partial X_R}$ and differentials $dY^{k\ell}{}_r$ leads to

$$dY_R = \left(\frac{\partial Y_R}{\partial X_R}\right)^T dX_R \ . \qquad (3.10.3)$$

Comparing $dY_R = PdX_R$ with (3.10.3), we obtain

$$\frac{\partial Y_R}{\partial X_R} = P^T \ ,$$

which proves (3.10.1).

(ii) is obvious from Theorem 2.10.1.

Here we notice that Theorem 3.10.1 is an extension of Theorem 3.4.1 to partitioned matrix functions.

The following theorem is an extension of Theorem 3.4.2 to partitioned situations.

Theorem 3.10.2 Let $Y = F(X)$ be any differentiable partitioned matrix function of $X{:}p{\times}q$, where X is partitioned into m row blocks and n column blocks. Suppose further that A_i, B_i, C_j, D_j ($i = 1, 2, \ldots, s$; $j = 1, 2, \ldots, t$) are some appropriate partitioned matrix functions of X. Then

$$dY = \sum_{i=1}^{s} A_i(dX)B_i + \sum_{j=1}^{t} C_j(dX^T)D_j \qquad (3.10.4)$$

and

$$\frac{\partial Y_R}{\partial X_R} = \sum_{i=1}^{s} (A_i{}^T \circledast B_i) + \sum_{j=1}^{t} (D_j \circledast C_j{}^T)I_{\{n\}} \qquad (3.10.5)$$

Proof: We proceed as in Theorem 3.4.2 and apply property (2.7.6), Theorems 2.10.1, 3.10.1 and expressions (2.8.9) and (2.9.2).

Here we observe the applications of the auxiliary matrices ${}_{\{m\}}I$, $I_{\{n\}}$ and the matrix product ⓟ in expressing the required matrix derivatives in a very compact form. One reason for partitioned matrix differentiation to be not considered so far seems to be the algebraic complications involved in partitioned matrix operations. Some of these difficulties are removed by introducing certain new ideas, for example, partitioned auxiliary matrices and the matrix product ⓟ .

In partitioned matrix differentiation, ${}_{\{m\}}I$ ($I_{\{n\}}$) plays the same role as $I_{(m)}$ ($I_{(n)}$) in non-partitioned matrix differentiation. Hence using ${}_{\{m\}}I$, we obtain expressions for $\frac{\partial Y_R}{\partial X_R}$ symbolically from non-partitioned results by Tracy & Dwyer (1969, Table 3) on using the transformation

$$\left(\frac{\partial Y_r}{\partial X_r}\ ,\ \otimes,\ I_{(m)}\ ,\ I_m\ ,\ I_n\right) \longrightarrow \left(\frac{\partial Y_R}{\partial X_R}\ ,\ ⓟ\ ,\ {}_{\{m\}}I,\ I_p\ ,\ I_q\right).$$

Note, however, that X here is a p×q matrix, with m row blocks and n column blocks, rather than X:m×n as in the paper of Tracy & Dwyer (1969).

Some results for partitioned matrix derivative $\frac{\partial Y_R}{\partial X_R}$ are given in the following table.

TABLE 3.10.1 PARTITIONED MATRIX DERIVATIVE $\frac{\partial Y_R}{\partial X_R}$ FOR SOME MATRIX FUNCTIONS

Y	$\frac{\partial Y_R}{\partial X_R}$
AX	$A^T \circledast I_q$
AX^T	$_{\{m\}}I(A^T \circledast I_p)$
XB	$I_p \circledast B$
X^TB	$_{\{m\}}I(I_q \circledast B)$
AXB	$A^T \circledast B$
AX^TB	$_{\{m\}}I(A^T \circledast B)$
X^TX	$_{\{m\}}I(I_q \circledast X) + X \circledast I_q$
X^TAX	$_{\{m\}}I(I_q \circledast AX) + A^TX \circledast I_q$
XX^T	$I_p \circledast X^T + {}_{\{m\}}I(X^T \circledast I_p)$
XBX^T	$I_p \circledast BX^T + {}_{\{m\}}I(B^TX^T \circledast I_p)$
CX^TAXD	$_{\{m\}}I(C^T \circledast AXD) + A^TXC^T \circledast D$
$CXBX^TD$	$C^T \circledast BX^TD + {}_{\{m\}}I(B^TX^TC \circledast D)$
X^{-1}	$-X^{-T} \circledast X^{-1}$
$AX^{-1}B$	$-X^{-T}A^T \circledast X^{-1}B$
$(AX^TB)^{-1}$	$-{}_{\{m\}}I\left[A^T(B^TXA^T)^{-1} \circledast B(AX^TB)^{-1}\right]$

Note: $\circledast$ denotes the circled-π (partitioned product) operator printed in the original.

TABLE 3.10.1 PARTITIONED MATRIX DERIVATIVE $\frac{\partial Y_R}{\partial X_R}$ FOR SOME MATRIX FUNCTIONS-Continued

Y	$\frac{\partial Y_R}{\partial X_R}$
$C(AXB)^{-1}D$	$-A^T(AXB)^{-T}C^T \circledast B(AXB)^{-1}D$
$XAX^{-1}BX^T$	$I_p \circledast AX^{-1}BX^T - (X^{-T}A^TX^T \circledast X^{-1}BX^T)$ $+ {}_{\{m\}}I(B^TX^{-T}A^TX^T \circledast I_p)$
$X^TAX^{-T}BX$	${}_{\{m\}}I(I_p \circledast AX^{-T}BX) -$ ${}_{\{m\}}I(X^{-1}A^TX \circledast X^{-T}BX) + (B^TX^{-1}A^TX \circledast I_p)$
X^α, α positive integer	$\sum_{i=0}^{\alpha-1}(X^T)^i \circledast X^{\alpha-i-1}$
$(X^T)^\alpha$, α positive integer	${}_{\{m\}}I\sum_{i=0}^{\alpha-1}\left[X^i \circledast (X^T)^{\alpha-i-1}\right]$
X^α, α negative integer	$-\sum_{i=1}^{-\alpha}(X^T)^{-i} \circledast X^{\alpha+i-1}$
$(X^T)^\alpha$, α negative integer	$-{}_{\{m\}}I\sum_{i=1}^{-\alpha}\left[X^{-i} \circledast (X^T)^{\alpha+i-1}\right]$

Results for $\frac{\partial Y_R}{\partial X_C}$, $\frac{\partial Y_C}{\partial X_R}$ and $\frac{\partial Y_C}{\partial X_C}$ are very easily obtained from $\frac{\partial Y_R}{\partial X_R}$ by using Theorem 2.10.1. These may also be obtained from the corresponding non-partitioned matrix derivative formulae by using the following transformation:

$(Y_r, Y_c, \otimes, I_{(m)}, I_{(n)}, I_m, I_n) \rightarrow (Y_R, Y_C, \textcircled{\pi}, {}_{\{m\}}I, I_{\{n\}}, I_p, I_q)$.

The partitioned auxiliary matrix ${}_{\{m\}}I$ uses the partitioning scheme $p_1q_1 \ldots p_mq_1 \ldots \quad \ldots p_1q_n \ldots p_mq_n$ for differentiating Y_R (Y_C) with respect to X_R, whereas for differentiating Y_R (Y_C) with respect to X_C, the partitioned identity matrix $I_{\{n\}}$ uses $p_1q_1 \ldots p_1q_n \ldots \quad \ldots p_mq_1 \ldots p_mq_n$ as its partitioning scheme.

In the following table, the non-partitioned matrix derivative results of Tracy & Dwyer (1969, Table 2) are extended to the partitioned situation:

TABLE 3.10.2 FORMULAE FOR $\frac{\partial Y_R}{\partial X_R}$, $\frac{\partial Y_R}{\partial X_C}$, $\frac{\partial Y_C}{\partial X_R}$, $\frac{\partial Y_C}{\partial X_C}$

WHEN $dY = A(dX)B$ or $C(dX^T)D$

	$dY = A(dX)B$		$dY = C(dX^T)D$	
Partial of → / w.r.t. ↓	Y_R	Y_C	Y_R	Y_C
X_R	$A^T \textcircled{\pi} B$	${}_{\{m\}}I(B \textcircled{\pi} A^T)$	${}_{\{m\}}I(C^T \textcircled{\pi} D)$	$D \textcircled{\pi} C^T$
X_C	$(I_{\{n\}})(A^T \textcircled{\pi} B)$	$B \textcircled{\pi} A^T$	$C^T \textcircled{\pi} D$	$(I_{\{n\}})(D \textcircled{\pi} C^T)$

From the above table, formula (3.10.5) for $\frac{\partial Y_R}{\partial X_R}$ is easy to derive for a linear combination of the form (3.10.4). Similarly if we require $\frac{\partial Y_C}{\partial X_R}$ from (3.10.4), then Table 3.10.2 yields:

$$\frac{\partial Y_C}{\partial X_R} = \sum_{i=1}^{s} ({}_{\{m\}}I)(B_i \circledR A_i^T) + \sum_{j=1}^{t} (D_j \circledR C_j^T).$$

Expressions for $\frac{\partial Y_R}{\partial X_C}$ and $\frac{\partial Y_C}{\partial X_C}$ are similarly obtained in Table 3.10.3.

Since

$$|I_{(m)}| = |I_{(n)}| = |{}_{\{m\}}I| = |I_{\{n\}}| = \pm 1,$$

the jacobian of a certain matrix transformation is the same no matter whether one uses $\frac{\partial Y_r}{\partial X_r}, \ldots, \frac{\partial Y_c}{\partial X_c}, \frac{\partial Y_R}{\partial X_R}, \ldots$ or $\frac{\partial Y_C}{\partial X_C}$.

Analogously, the concepts of Table 3.4.1 and Figure 3.7.1 may also be extended to the partitioned situation.

TABLE 3.10.3 FORMULAE FOR $\frac{\partial Y_R}{\partial X_R}$, $\frac{\partial Y_R}{\partial X_C}$, $\frac{\partial Y_C}{\partial X_R}$, $\frac{\partial Y_C}{\partial X_C}$ WHEN

$$dY = \sum_i A_i(dX)B_i + \sum_j C_j(dX^T)D_j$$

Partial of → / w.r.t. ↓	Y_R	Y_C
X_R	$\sum_i(A_i^T \textcircled{\pi} B_i)+({}_{\{m\}}I)\sum_j(C_j^T \textcircled{\pi} D_j)$	$({}_{\{m\}}I)\sum_i(B_i \textcircled{\pi} A_i^T)+\sum_j(D_j \textcircled{\pi} C_j^T)$
X_C	$(I_{\{n\}})\sum_i(A_i^T \textcircled{\pi} B_i)+\sum_j(C_j^T \textcircled{\pi} D_j)$	$\sum_i(B_i \textcircled{\pi} A_i^T)+(I_{\{n\}})\sum_j(D_j \textcircled{\pi} C_j^T)$

3.11 Differential of Functions of Matrices with Equality Relationships Among Their Scalar Elements

The theory developed in the above sections of this chapter, which we recommend for general use, calls for the differentiation of each matrix element of Y with respect to the matrix elements of X, where Y = F(X) is a matrix function. The concept of a matrix element versus a scalar element of a matrix is discussed in Section 3.1. From this discussion it is clear that the matrix elements are not completely identified by their scalar values alone since their position in the matrix must be specified as well. Thus it may happen that $x_{ij} = x_{k\ell}$, but always $\langle x_{ij} \rangle \neq \langle x_{k\ell} \rangle$, when $i \neq k$ or $j \neq \ell$ or both.

It seems then appropriate to define matrix derivatives with respect to the matrix elements themselves and not their scalar values. Thus, with the use of $\frac{\partial Y_r}{\partial X_r}$, $\frac{\partial Y_r}{\partial X_c}$, $\frac{\partial Y_c}{\partial X_r}$, and $\frac{\partial Y_c}{\partial X_c}$, the elements of the matrix derivatives are defined to be the elements of the matrices (scalar values in position) and not just the scalar values. Several general results concerning the above definition of matrix derivatives are given in the previous sections of this chapter.

There are various kinds of relationships which may exist between the scalar values of the matrix elements. However, in our development, we say two scalar elements x_{ij} and $x_{k\ell}$ of X have known equality relationships if $x_{ij} = x_{k\ell}$ where $i \neq k$ or

$j \neq \ell$ or both. In matrix X, there may be several such equality relationships between pairs of elements; however no three elements should be equal to each other. The matrix derivative theory so developed is useful in obtaining jacobians of the symmetric matrix transformations, where $x_{ij} = x_{ji}$ for $i \neq j$, and no other scalar elements are equal. No general theory seems feasible to cover various combinations of possible equality relationships. For further development and applications of this concept (Section 4.7), only this restricted set of "known equality relationships" is treated. In particular, a subset of the most commonly known equality relationships in many statistical problems is obtained from the concept of symmetry. For example, in evaluating beta and gamma integrals for symmetric matrices, we come across the symmetric matrix transformations

$$Z = A^T A$$

and

$$Y = AXA^T$$

where A, Z, X and Y are positive definite symmetric matrices, (see Jack (1964-65, pp. 97-98)). In general we may have matrix functions $Y = F(X, Z)$ with arbitrary scalar relationships among matrix elements of Y which may be due to arbitrary equality relationships among the scalar values of its independent matrix variables. An application of the concept of matrix element of a matrix to the theory of matrix differential calculus is pointed out by Dwyer (1970). In a later paper, Tracy & Singh (1970a) presented some formulae concerning differentiation of functions of matrices involving a few known equality relationships among scalar elements and applied it to a statistical problem. Most of the results of this section are repeated from the above paper for completeness.

Using the above <u>matrix</u> definition of the collection of partial derivatives, we can obtain the matrix derivatives with respect to the matrix elements of the matrix X, since $\langle x_{ij} \rangle$ is always independent of $\langle x_{i'j'} \rangle$ except when $i' = i$, $j' = j$, even though relationships may be known to exist between the scalar values of the elements. These derivatives are unaffected even if there are known or unknown equality relationships among the scalar values of matrix elements. In this section, we consider modifications of the following two types of matrix derivative of matrix function $Y = F(X)$ involving <u>known equality relationships</u> among the scalar values of the matrix elements:

(i) $\frac{\partial Y_r}{\partial X_r}$, which is obtained by differentiating each vector element of Y_r with respect to the vector elements of X_r,

(ii) the matrix derivative, obtained by differentiating each scalar element of Y_r with respect to the scalar elements of X_r.

In addition, we obtain matrix derivative described in (ii) as a linear matrix function of $\frac{\partial Y_r}{\partial X_r}$.

The above modifications lead to throwing out certain repeated rows and columns of the matrix derivatives. Although, in throwing out repeated rows and columns, we have changed the problem, still the corresponding determinantal values yield jacobians of the original problem. We accept this fact simply because we have thrown out only those rows and columns which did not reduce the number of distinct rows and columns in the modified matrix derivatives (see Tracy & Dwyer (1969, pp. 1582-1583)).

The auxiliary matrices $I_{(k)}$ and the non-auxiliary matrices M, N and I* defined in Section 2.4 are used in identifying the above mentioned matrix derivatives.

The modified versions of these matrix derivatives are used in Section 4.7 to evaluate the jacobians of certain symmetric matrix transformations.

3.11.1 Some Definitions and Interrelationships

Here we introduce some column vectors and matrices which are certain deformations of X_r and the identity matrices.

Definition 3.11.1.1 Let the matrix X have some known equality relationships among the scalar values of its matrix elements. Then we obtain the column vector $X_r^{\#}$ by cancelling the matrix elements corresponding to repeated scalar elements as we go down the vector X_r. This defines $X_r^{\#}$ uniquely.

Definition 3.11.1.2 Let X be as in the above definition. Then we define $X_{(r)}$ as a column vector of the scalar elements of the corresponding matrix elements of X_r; and $X_{(r)}^{\#}$ as a column vector of the distinct scalar elements of $X_{(r)}$.

Here we observe that the elements of $X_{(r)}$ and $X_{(r)}^{\#}$ are not unique. However, we agree that a subset of distinct elements chosen from $X_{(r)}$ to obtain $X_{(r)}^{\#}$, though not necessarily unique, behave as independent variables. In the above definitions, the repeated scalar values of matrix elements can be permuted among themselves and hence permit algebraic simplification with the use of known equality relationships. We denote these elements arranged in vector form $X_{(r)}$ by $\{X_{ij}\}$.

With reference to a matrix variable X and certain identity matrices, we define the matrices M_X , N_X and I_X^* .

Let $X{:}m{\times}n$ be such that k of the scalar value of its matrix elements are repeated. Let these be identified as the i-th elements of $X_{(r)}$, for k values of $i \in \{2,\ldots,mn\}$.

Then

(1) a matrix obtained by deleting k, i-th rows from $I{:}mn{\times}mn$ is denoted by $M_X{:}(mn-k){\times}mn$,

(2) a matrix, obtained by adding to $I{:}(mn-k){\times}(mn-k)$ k, i-th rows corresponding to the repeated elements of $X_{(r)}$ which have occurred before, is denoted by $N_X{:}mn{\times}(mn-k)$,

(3) a diagonal matrix with 1 corresponding to elements of $X_{(r)}$ involving equality relationships and 0 elsewhere is denoted by $I_X^*{:}mn{\times}mn$.

The properties of these matrices are the same as those of M, N and I^* given in Theorem 2.4.2.

Some useful equalities involving the above concepts are given by the following

<u>Theorem 3.11.1.1</u> For any matrix X with some known equality relationships and no others, we have

$$X_r^{\#} = M_X X_r \tag{3.11.1.1}$$

$$X_{(r)}^{\#} = M_X X_{(r)} \tag{3.11.1.2}$$

$$X_{(r)} = N_X X_{(r)}^{\#} , \tag{3.11.1.3}$$

where the matrices and the vectors involved are conformable.

This theorem may be easily verified.

The following examples illustrate the above definitions and Theorem 3.11.1.1.

<u>Example 3.11.1.1</u> Let X:2×3 matrix such that $x_{13} = x_{21}$. Then

$$X_r = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{23} \end{bmatrix}, \quad X_r^{\#} = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{22} \\ x_{23} \end{bmatrix}, \quad X_{(r)} = \begin{Bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{23} \end{Bmatrix}, \quad X_{(r)}^{\#} = \begin{Bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{22} \\ x_{23} \end{Bmatrix},$$

$$M_X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad N_X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad I_X^{*} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and Theorem 3.11.1.1 holds.

<u>Example 3.11.1.2</u> For symmetric matrices the situation becomes very straightforward. Let $X = X^T$:3×3, then

$$X_r = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{31} \\ x_{32} \\ x_{33} \end{bmatrix}, \quad X_r^{\#} = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{22} \\ x_{23} \\ x_{33} \end{bmatrix}, \quad X_{(r)} = \left\{ \begin{matrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{31} \\ x_{32} \\ x_{33} \end{matrix} \right\}, \quad X_{(r)}^{\#} = \left\{ \begin{matrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{22} \\ x_{23} \\ x_{33} \end{matrix} \right\},$$

$$M_X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad N_X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$I_X^{*} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The results of Theorem 3.11.1.1 are easily verified.

Expressions (3.11.1.2) and (3.11.1.3) yield the following:

Theorem 3.11.1.2 If $Y_{(r)} = KX_{(r)}$ such that the scalar elements of $Y_{(r)}$ and $X_{(r)}$ have known equality relationships among themselves, then

$$Y_{(r)}^{\#} = M_Y K N_X X_{(r)}^{\#} . \qquad (3.11.1.4)$$

In the following discussion we give explicit meaning to the matrices M_Y and N_X which occur in the proof of certain theorems involving symmetric matrices. Here we remark that the matrices M_Y , N_X and I_X^* may be of more general nature, depending on the type of known relationships (for example, some functional relationships), among scalar elements of Y and X respectively. The only modification needed to obtain M_Y and N_X is to delete and add appropriate types of rows in the respective matrices described above. A modification to obtain I_X^* may also be worked out. However, there does not exist any general procedure which takes care of all known relationships.

In obtaining the jacobians of certain matrix transformations, it is desirable to know the distinct scalar elements of both Y and X for a matrix function Y = F(X). The determinant of the matrix derivative of each of the distinct scalar elements of Y with respect to the distinct scalar elements of X is the required jacobian except for sign. For matrix functions involving known equality relationships among their scalar elements, we obtain the above mentioned matrix derivative and some other results in the following subsections.

As far as applications are concerned, we require Corollary 3.11.2.1 and Theorem 3.11.3.3, since in evaluating certain multiple integrals we come across symmetric matrix transformations.

3.11.2 Modification of Matrix Derivatives with Respect to Matrix Elements

The fact that

$$d\langle x_{ij}\rangle = \langle dx_{ij}\rangle = (i,j,dx_{ij}) \qquad (3.11.2.1)$$

is useful in expressing differential dX_r as a function of $dX_{(r)}^{\#}$. Since any two different matrix elements are always distinct, for any non-singular matrix transformation $Y = F(X)$, the determinant of $\frac{\partial Y_r}{\partial X_r}$ is non-zero. However, if any two different matrix elements have equal scalar values, then the absolute value of the determinant of $\frac{\partial Y_r}{\partial X_r}$ is not the appropriate jacobian. Hence we modify $\frac{\partial Y_r}{\partial X_r}$ to obtain the appropriate jacobian. Above results are discussed in the following:

Theorem 3.11.2.1 For any matrix function $Y = F(X)$, with known equality relationships among scalar elements of $Y:m\times n$, $X:p\times q$ themselves, we have

(i) $dX_r = N_X dX_{(r)}^{\#}$ (3.11.2.2)

(ii) $dY_r = H dX_r \Longrightarrow \dfrac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}} = (M_Y H N_X)^T$, (3.11.2.3)

where M_Y and N_X are (mn-k)×mn, pq×(pq-ℓ) matrices obtained by appropriately deleting k rows from or adding ℓ rows to identity matrices of orders mn, pq-ℓ (k, ℓ are the number of known equality relationships in Y, X) respectively.

Proof: (i) We claim that even though $\langle x_{ij} \rangle$ and x_{ij} are not completely independent variables, their differentials are equal. Thus $d\langle x_{ij} \rangle = dx_{ij}$ in the (i,j)-th position and (3.11.2.2) holds.

(ii) We have

$$dY_r^{\#} = M_Y H dX_r$$

$$= M_Y H N_X dX_{(r)}^{\#} \qquad (3.11.2.4)$$

and

$$dY_r^{\#} = \left(\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}}\right)^T dX_{(r)}^{\#} . \qquad (3.11.2.5)$$

Comparison of (3.11.2.4) and (3.11.2.5) yields (3.11.2.3).

The following examples illustrate (3.11.2.3).

Example 3.11.2.1 Let $y = \underline{a}^T X \underline{b}$ be a scalar differentiable function of X; $\underline{a}$, $\underline{b}$ are column vectors, X has some known equality relationships, then

$$dy = (\underline{a}^T \otimes \underline{b}^T) dX_r$$

$$= (\underline{a}^T \otimes \underline{b}^T) N_X dX_{(r)}^{\#} .$$

Hence

$$\frac{\partial y}{\partial X_{(r)}^{\#}} = N_X^T(\underline{a}\otimes\underline{b}) .$$

Example 3.11.2.2 Let $Y = XAX^T$ be a matrix function, where $X{:}p\times q$, and X and Y have some known equality relationships, then

$$dY = (dX)AX^T+XA(dX^T)$$

$$\Longrightarrow \quad dY_r = (I\otimes XA^T)dX_r+(XA\otimes I)(dX^T)_r$$

$$\Longrightarrow \quad dY_r^{\#} = M_Y[(I\otimes XA^T)+(XA\otimes I)I_{(q)}]N_X dX_{(r)}^{\#}$$

$$\Longrightarrow \quad \frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}} = N_X^T[(I\otimes AX^T)+I_{(p)}(A^TX^T\otimes I)]M_Y^T .$$

The following corollary has applications in multivariate analysis.

Corollary 3.11.2.1 If, in Theorem 3.11.2.1, $Y = Y^T{:}m\times m$, $X = X^T{:}p\times p$, then

$$\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}} = [M_Y H N_X]^T , \qquad (3.11.2.6)$$

where M_Y and N_X are $\frac{m(m+1)}{2}\times m^2$ and $p^2\times\frac{p(p+1)}{2}$ appropriate matrices respectively as before.

The matrix derivative $\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}}$ is useful in evaluating the jacobian of certain mutilated matrix transformations. Using this matrix derivative, some important statistical applications of the above corollary are given in Section 4.7.

3.11.3 Modification of Matrix Derivatives with Respect to Scalar Elements

The statistical applicability of this section is the same as that of Section 3.11.2. In this section we establish an expression which verifies an important fact pointed out by Dwyer (1970). Here we require the following definition of matrix derivative with respect to scalar elements.

Definition 3.11.3.1 For any matrix function $Y = F(X)$, we define $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$ to be a matrix of partial derivatives of scalar elements y_{ij} with respect to the scalars $x_{k\ell}$, uniquely ordered.

We denote the matrix derivative $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$ by $\left\{\frac{\partial y_{ij}}{\partial x_{k\ell}}\right\}$.

A procedure for identifying $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$ is as follows:

Theorem 3.11.3.1 For any matrix function $Y = F(X)$, we have

$$dY_{(r)} = KdX_{(r)} \Longrightarrow \frac{\partial Y_{(r)}}{\partial X_{(r)}} = K^T . \qquad (3.11.3.1)$$

There exists a natural relationship between the matrix derivatives $\frac{\partial Y_r}{\partial X_r}$ and $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$. We discuss a particular situation in the following:

Theorem 3.11.3.2 For any matrix function $Y = F(X)$, where the scalar elements of Y may or may not have equality relationships, and where $X = X^T$:$p\times p$ are the only known equality relationships among the scalar values of $<x_{ij}>$ for $i \neq j$, we have

$$\frac{\partial Y_{(r)}}{\partial X_{(r)}} = \frac{\partial Y_r}{\partial X_r} + I^*{}_{(p)}\frac{\partial Y_r}{\partial X_r} , \tag{3.11.3.2}$$

where I^* is a $p^2 \times p^2$ diagonal matrix with zeros in positions 1 mod(p+1) and 1's elsewhere.

Proof: We see that

$$dy_{k\ell} = \Bigg\{ \ldots ; \frac{\partial <y_{k\ell}>}{\partial <x_{i1}>} + \frac{\partial <y_{k\ell}>}{\partial <x_{1i}>} \ldots \frac{\partial <y_{k\ell}>}{\partial <x_{i,i-1}>} + \frac{\partial <y_{k\ell}>}{\partial <x_{i-1,i}>} ,$$

$$\frac{\partial <y_{k\ell}>}{\partial <x_{ii}>} , \frac{\partial <y_{k\ell}>}{\partial <x_{i+1,i}>} + \frac{\partial <y_{k\ell}>}{\partial <x_{i,i+1}>} \ldots \frac{\partial <y_{k\ell}>}{\partial <x_{ip}>} +$$

$$\frac{\partial <y_{k\ell}>}{\partial <x_{pi}>} ; \ldots \Bigg\} dX_{(r)} \quad . \tag{3.11.3.3}$$

Hence from (3.11.3.3), we obtain

$$dY_{(r)} = \left[\frac{\partial Y_r}{\partial X_r} + I^* \; I_{(p)} \; \frac{\partial Y_r}{\partial X_r}\right]^T dX_{(r)} \tag{3.11.3.4}$$

The required result follows from (3.11.3.1) and (3.11.3.4).

Example 3.11.3.1 Let $Y = y$ and $X = X^T : 2\times 2$, then

$$\frac{\partial Y_{(r)}}{\partial X_{(r)}} = \begin{Bmatrix} \frac{\partial y}{\partial x_{11}} \\ \frac{\partial y}{\partial x_{12}} \\ \frac{\partial y}{\partial x_{21}} \\ \frac{\partial y}{\partial x_{22}} \end{Bmatrix} = \begin{Bmatrix} \frac{\partial y}{\partial \langle x_{11} \rangle} \\ \frac{\partial y}{\partial \langle x_{12} \rangle} + \frac{\partial y}{\partial \langle x_{21} \rangle} \\ \frac{\partial y}{\partial \langle x_{21} \rangle} + \frac{\partial y}{\partial \langle x_{12} \rangle} \\ \frac{\partial y}{\partial \langle x_{22} \rangle} \end{Bmatrix}$$

$$= \begin{bmatrix} \frac{\partial y}{\partial \langle x_{11} \rangle} \\ \frac{\partial y}{\partial \langle x_{12} \rangle} \\ \frac{\partial y}{\partial \langle x_{21} \rangle} \\ \frac{\partial y}{\partial \langle x_{22} \rangle} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial y}{\partial \langle x_{11} \rangle} \\ \frac{\partial y}{\partial \langle x_{12} \rangle} \\ \frac{\partial y}{\partial \langle x_{21} \rangle} \\ \frac{\partial y}{\partial \langle x_{22} \rangle} \end{bmatrix}$$

$$= \frac{\partial y}{\partial X_r} + I^*_{(2)} \frac{\partial y}{\partial X_r} \quad .$$

Hence (3.11.3.2) is valid for this particular example.

Theorem 3.11.3.2 may be extended individually for more general situations. However, it is not possible to establish a general formula relating $\frac{\partial Y_r}{\partial X_r}$ and $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$ for a general matrix function involving arbitrary relationships, known or unknown, among the scalar elements of Y and of X. For example, $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$ has different expressions for X: symmetric matrix and

X: non-symmetric matrix, whereas $\frac{\partial Y_r}{\partial X_r}$ remains unchanged.

We note that if any two matrix elements have equal scalar values, then $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$ becomes singular and the jacobian of the transformation cannot be obtained unless we get rid of those rows and columns which are repeated. This leads to the following modification of Theorems 3.11.3.1 and 3.11.3.2:

Theorem 3.11.3.3 For any matrix function Y = F(X) where the scalar elements of Y and X have only known equality relationships among themselves, we obtain

$$dY_{(r)} = KdX_{(r)} \Longrightarrow \frac{\partial Y_{(r)}{}^{\#}}{\partial X_{(r)}{}^{\#}} = \left[M_Y K M_X{}^T\right]^T . \qquad (3.11.3.5)$$

Proof: Since we need to remove the repeated rows and repeated columns from $\frac{\partial Y_{(r)}}{\partial X_{(r)}}$, we pre-multiply and post-multiply the matrix K^T by appropriate M_X and $M_Y{}^T$ to get $\frac{\partial Y_{(r)}{}^{\#}}{\partial X_{(r)}{}^{\#}}$.

The following is an example illustrating the above theorems.

Example 3.11.3.2 If Y = AXB, where A and B are constant matrices of conformable orders, $X = X^T\!:\!p{\times}p$, then

$$\frac{\partial Y_{(r)}}{\partial X_{(r)}{}^{\#}} = M_X[A^T \otimes B + I^*{}_{(p)}(A^T \otimes B)] . \qquad (3.11.3.6)$$

If, in addition, $Y = Y^T\!:\!m{\times}m$, then

$$\frac{\partial Y_{(r)}{}^{\#}}{\partial X_{(r)}{}^{\#}} = M_X[A^T \otimes B + I^*{}_{(p)}(A^T \otimes B)]M_Y{}^T \ , \qquad (3.11.3.7)$$

where I^* is as in Theorem 3.11.3.2 and M_X , M_Y are suitable non-auxiliary matrices.

From Theorem 3.11.3.2, we observe that <u>in general</u>:

$$\frac{\partial Y_{(r)}}{\partial X_{(r)}} \neq \frac{\partial Y_r}{\partial X_r} \qquad (3.11.3.8)$$

From Theorems 3.11.2.1 and 3.11.3.3, we can verify that, <u>in general</u>:

$$\frac{\partial Y_{(r)}{}^{\#}}{\partial X_{(r)}{}^{\#}} = \frac{\partial Y_r{}^{\#}}{\partial X_{(r)}{}^{\#}} \ . \qquad (3.11.3.9)$$

If, in Theorem 3.11.3.2, X is neither a symmetric matrix nor any of its scalar elements have equality relationships, then $I^*{}_{(p)} = 0$. Hence (3.11.3.2) reduces to

$$\frac{\partial Y_{(r)}}{\partial X_{(r)}} = \frac{\partial Y_r}{\partial X_r} \ . \qquad (3.11.3.10)$$

CHAPTER IV

APPLICATIONS TO GENERAL MULTIVARIATE MODELS

4.1 Introduction

This chapter seeks to focus attention on some applications of the matrix differential calculus presented in Chapter III to general multivariate statistical models. Dwyer (1958, 1967), Neudecker (1967, 1968, 1969b) and Tracy & Dwyer (1969) have applied their theoretical results on matrix differentiation to

1. obtain maximum likelihood estimates for multivariate models,
2. consider constrained optimization of matrix functions,
3. evaluate jacobians of certain matrix transformations applicable in multivariate distribution theory.

Stroud (1968), in obtaining asymptotic tests for (1) equality of conditional covariance matrices and (2) equality of conditional mean vectors when in both cases, errors of measurement have known variances, used results of Dwyer & MacPhail (1948) and Dwyer (1967). Anderson (1968) considered maximum likelihood estimation of coefficients when the covariance matrix has linear structure, without using matrix differentiation. Some important applications of Dwyer's (1967) matrix differentiation results have recently been pointed out by Kleinbaum (1970). Kleinbaum discusses estimation and testing hypotheses for multivariate linear models in which some observations are missing and/or in which different design matrices correspond to different response variates. Further applications of less general results on matrix differentiation have been considered by various authors: Dwyer (1958, 1967), Goldberger (1964), Trawinski & Bargmann (1964), Rao (1965), Bock & Bargmann (1966), Eisenpress & Greenstadt (1966), Srivastava (1966), Jöreskog (1966, 1967, 1969, 1970a, 1971), Jöreskog & Lawley (1968), Fisk (1967), Lawley (1967), Morrison (1967), Tan (1968-69), Gebhardt (1971) and Mulaik (1971).

Multivariate problems discussed in Dwyer (1958, 1967), Fisk (1967), Neudecker (1967, 1968, 1969b), Tracy & Dwyer (1969), Gebhardt (1971) and Mulaik (1971) show various applications of matrix differentiation formulae. These formulae have been established in the above papers, though, some of them have followed different approaches.

In the present chapter, we reformulate some of the models in a more general setting. Some of these models are of econometric and psychometric interest. Matrix differentiation results of Chapter II are then useful in the estimation of parameters, for finding asymptotic covariance matrices and for evaluating the jacobians of symmetric matrix transformations. These results are also applicable in a k-sample regression analysis with covariance, in the analysis of covariance matrices with linear structure and in the dynamic econometric analysis. An auxiliary matrix from Section 2.4 and its extension are used in obtaining some properties of the matrix products $\odot$ and $\textcircled{e}$.

4.2 Estimation of Parameters in General Multivariate Linear Regression Analysis

Consider the regression model of observations X

$$X = A\Xi P+U \tag{4.2.1}$$

where A:n×g and P:h×p are model matrix and fixed matrix respectively, Ξ:g×h is a matrix of parameters to be estimated, and U:n×p is a matrix of random errors. We assume that

$$E(U) = 0 \tag{4.2.2}$$

and

$$\mathrm{Var}(U_r) = V\otimes\Sigma \tag{4.2.3}$$

where V:n×n and Σ:p×p are positive definite covariance matrices. The general method of least squares and maximum likelihood principle can be applied to the model (4.2.1) for estimating parameters Ξ, V and Σ.

4.2.1 Least Squares Estimation

A least squares estimator of the parameter matrix Ξ is $\hat{\Xi}$ obtained by minimizing the function

$$f(\Xi) = \mathrm{tr}\Sigma^{-1}(X-A\Xi P)^TV^{-1}(X-A\Xi P). \tag{4.2.1.1}$$

We apply Theorem 3.4.3 and Corollary 3.8.2 to obtain first and second order matrix derivatives of f(Ξ) with respect to the unknown regression parameter matrix Ξ.

On differentiating f(Ξ), we obtain

$$df(\Xi) = \mathrm{tr}[-2\Sigma^{-1}X^TV^{-1}A(d\Xi)P+\Sigma^{-1}P^T(d\Xi)^TA^TV^{-1}A\Xi P$$
$$+\Sigma^{-1}P^T\Xi^TA^TV^{-1}A(d\Xi)P]. \tag{4.2.1.2}$$

Using Theorem 3.4.3, we have

$$\frac{\partial f(\Xi)}{\partial \Xi} = -2A^T V^{-1}[X-A\Xi P]\Sigma^{-1}P^T \ , \tag{4.2.1.3}$$

which, on setting $\frac{\partial f(\Xi)}{\partial \Xi} = 0$, yields the normal equation

$$A^T V^{-1} X \Sigma^{-1} P^T = A^T V^{-1} A \hat{\Xi} P \Sigma^{-1} P^T \ . \tag{4.2.1.4}$$

To obtain the second order matrix derivative, we have, from expression (4.2.1.2),

$$d^2 f(\Xi) = 2\mathrm{tr}[\Sigma^{-1}P^T(d\Xi^T)A^T V^{-1}A(d\Xi)P] \ . \tag{4.2.1.5}$$

We have, on using Corollary 3.8.2,

$$\nabla^2_{\Xi_r,\Xi_r} f(\Xi) = A^T V^{-1} A \otimes P\Sigma^{-1}P^T \ . \tag{4.2.1.6}$$

If we assume that $A^T V^{-1} A$ and $P\Sigma^{-1}P^T$ are positive definite matrices, then from (4.2.1.4), for given V and Σ,

$$\hat{\Xi} = (A^T V^{-1} A)^{-1}(A^T V^{-1} X \Sigma^{-1} P^T)(P\Sigma^{-1}P^T)^{-1} \ , \tag{4.2.1.7}$$

which is the least squares estimator of Ξ, and $\nabla^2_{\Xi_r,\Xi_r} f(\Xi)$ is positive definite (cf. A.1.35). This proves that $f(\Xi)$ attains its minimum at $\hat{\Xi}$, on using Theorem 3.9.1.

Tan (1968-69) obtained the normal equations for the model (4.2.1) with P = I, by using a very complicated method. Also he does not mention sufficient conditions for the minimum.

The general linear model (4.2.1) becomes simpler if P = I, and $\mathrm{Var}(U_r) = I$. This case has been considered by Neudecker (1967) and Tracy & Dwyer (1969), using different approaches. They have obtained both necessary and sufficient conditions for minimizing

$$\log|(X-A\Xi)^T(X-A\Xi)| \ . \qquad (4.2.1.8)$$

Model (4.2.1) can be very easily considered by minimizing

$$\log|\Sigma^{-1}(X-A\Xi P)^T V^{-1}(X-A\Xi P)| \ , \qquad (4.2.1.9)$$

which yields the same results as are obtained by minimizing $f(\Xi)$ of (4.2.1.1).

4.2.2 Maximum Likelihood Estimation

We give a simpler proof of a lemma which is similar to a result in Anderson (1958, Lemma 3.2.2).

Lemma 4.2.2.1 A differentiable scalar function of X

$$f(X) = \frac{1}{q}[N \log|X| - \mathrm{tr}XD] \qquad (4.2.2.1)$$

attains its maximum at

$$\hat{X} = N\hat{D}^{-1} \ , \qquad (4.2.2.2)$$

where X and D are m×m positive definite matrices.

Proof: On differentiating,

$$df(X) = \frac{1}{q}[\mathrm{tr}NX^{-T}(dX^T) - \mathrm{tr}(dX)D] \ ,$$

which gives

$$\frac{\partial f(X)}{\partial X} = \frac{1}{q}[NX^{-T} - D^T] \ , \quad \text{using Theorem 3.4.3.}$$

Setting $\frac{\partial f(X)}{\partial X} = 0$ yields $X = ND^{-1}$.

Then using the same argument as in Anderson (1958, Lemma 3.2.2) $f(X)$ has its maximum at $\hat{X} = N\hat{D}^{-1}$. The procedure followed by Anderson (1958) in proving the above lemma is much more complicated than the one presented here.

The above lemma has an interesting application in multivariate regression analysis.

We suppose that the random error matrix U:n×p is normally distributed with mean 0 and covariance matrix $V\otimes\Sigma$. Then the likelihood function of the observation matrix X is given by

$$L(\Xi,V,\Sigma|X) = \pi^{np}|\Sigma|^{-\frac{n}{2}}|V|^{-\frac{p}{2}}e^{-\frac{1}{2}trV^{-1}(X-A\Xi P)\Sigma^{-1}(X-A\Xi P)^T} . \quad (4.2.2.3)$$

[Expression (4.1) of Tan (1968-69) is a particular case of (4.2.2.3) with $\Xi = I$.]

To find the maximum likelihood estimate of Ξ, we obtain a solution $\hat{\Xi}$ of the maximum likelihood equations

$$\frac{\partial \log L(\Xi,V,\Sigma|X)}{\partial\Xi} = 0 \quad (4.2.2.4)$$

The matrix function $\hat{\Xi}$ is the required maximum likelihood estimator.

Now treating (4.2.2.3) as a function of Ξ only, we get

$$d\log L(\Xi) = -\tfrac{1}{2}df(\Xi), \quad (4.2.2.5)$$

where $f(\Xi)$ is as in (4.2.1.1). From (4.2.2.4) and (4.2.2.5) we obtain the likelihood equation

$$A^TV^{-1}X\Sigma^{-1}P^T = A^TV^{-1}A\hat{\Xi}P\Sigma^{-1}P^T , \quad (4.2.2.6)$$

which is the same as (4.2.1.4). The second order matrix derivative

$$\nabla^2_{\Xi_r,\Xi_r}\log L(\Xi) = -\tfrac{1}{2}(A^TV^{-1}A\otimes P\Sigma^{-1}P^T) \quad (4.2.2.7)$$

can similarly be shown to be negative definite under the assumptions of the previous subsection. This shows that any solution $\hat{\Xi}$ of (4.2.2.6) is the maximum likelihood estimate.

Now suppose that V is a known covariance matrix. Then, as in Tan (1968-69, (4.2)) with P = I, we have

$$\log L(\Xi,\Sigma|X) = \text{constant}-\tfrac{n}{2}\log|\Sigma|-\tfrac{1}{2}\mathrm{tr}V^{-1}(X-A\Xi P)\Sigma^{-1}(X-A\Xi P)^T$$

$$= \text{constant}-\tfrac{n}{2}\log|\Sigma|-\tfrac{1}{2}\mathrm{tr}\Sigma^{-1}(X-A\Xi P)^T V^{-1}(X-A\Xi P) \quad (4.2.2.8)$$

Then from Lemma 4.2.2.1, $\log L(\Xi,\Sigma|X)$ attains its maximum at

$$\hat{\Sigma} = \tfrac{1}{n}(X-A\hat{\Xi}P)^T V^{-1}(X-A\hat{\Xi}P) \ , \quad (4.2.2.9)$$

which is the maximum likelihood estimate of Σ.

If, on the other hand, Σ is known, then proceeding exactly as above, the maximum likelihood estimate of V is given by

$$\hat{V} = \tfrac{1}{p}(X-A\hat{\Xi}P)\Sigma^{-1}(X-A\hat{\Xi}P)^T \ . \quad (4.2.2.10)$$

Here we emphasize the advantage of using Lemma 4.2.2.1 for drawing inferences regarding normally distributed random variables. This lemma identifies the maximum likelihood estimates of covariance matrices without formally using matrix differentiation. In the general least squares theory, we had to compute second order matrix derivatives because the functions to be minimized could not be transformed to functions of the type designated in Lemma 4.2.2.1, to be maximized.

Maximum likelihood estimates considered by Tracy & Dwyer (1969) are obtained using first and second order matrix derivatives. They have established the concavity of the log-likelihood function by showing that its Hessian matrix at the solutions is negative definite. The multivariate regression model considered by Tracy & Dwyer (1969) is obtained from the model (4.2.1), where $X = \underline{x}^T: 1\times p$, $A = \underline{z}^T: 1\times q$, $P = I_p$ and V=1.

4.3 Estimation of Parameters in Non-linear Multivariate Regression Analysis

Non-linear multivariate models have been considered by several authors following differerent approaches. Eisenpress & Greenstadt (1966) obtained expressions for estimating regression parameters of non-linear econometric systems. Allen (1967) considered parameter estimation, hypothesis testing and large sample properties of a general non-linear multivariate model which is useful in the analysis of growth curves. An extensive list of references on non-linear models is available in Allen (1967). The above authors discussed non-linear models using scalar differentiation.

In the non-linear multivariate models, each observation may be a non-linear function of the unknown parameter matrix. Consider

$$x_{ij} = f_i(\Xi, t_j)+u_{ij} \ , \qquad (4.3.1)$$

$$i = i, 2, \ldots, n; \ j = 1, 2, \ldots, p;$$

where the observations x_{ij} are non-linear functions of the elements of the parameter matrix Ξ:r×s. We note that $f_i(\Xi, t_j)$ are non-linear functions of the parameters $\beta_{\gamma\delta}$, where $\Xi=[\beta_{\gamma\delta}]$, the u_{ij} are random errors and the t_j are j-th points in time. A matrix representation of the model (4.3.1) is

$$X = AF(\Xi) + U \ , \qquad (4.3.2)$$

where X:n×p is the observation matrix; $F(\Xi)$:h×p is a matrix of non-linear functions of the elements of Ξ, A:n×h is a known matrix, and U:n×p is a matrix of random errors.

As in the general multivariate linear models, we assume that

$$E(U) = 0 \qquad (4.3.3)$$

and

$$\mathrm{Var}(U_r) = V \otimes \Sigma \ . \qquad (4.3.4)$$

If, in particular, $F(\Xi) = \Xi P$, then (4.3.2) is similar to the linear model considered in Section 4.2.

Example 4.3.1 Suppose there are two groups of 4 and 5 animals respectively, and each group has different parameter values, i.e.

$$f_i(\Xi, t_j) = \beta_{i1} + \beta_{i2}\beta_{i3}^{\,j-1} \ , \qquad (4.3.5)$$

and suppose that the response $f_i(\Xi, t_j)$ is measured for six days. Let $X_{k.}$, $k = 1, 2, \ldots, 4$ be the row vectors of six observations on each of the animals in the first group; and let $X_{\ell.}$, $\ell = 1, 2, \ldots, 5$ be observations on the second group. Then

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} ,$$

$$F(\Xi) = \begin{bmatrix} \beta_{11}+\beta_{12} & \beta_{11}+\beta_{12}\beta_{13} & \beta_{11}+\beta_{12}\beta_{13}^2 & \beta_{11}+\beta_{12}\beta_{13}^3 & \beta_{11}+\beta_{12}\beta_{13}^4 & \beta_{11}+\beta_{12}\beta_{13}^5 \\ \beta_{21}+\beta_{22} & \beta_{21}+\beta_{22}\beta_{23} & \beta_{21}+\beta_{22}\beta_{23}^2 & \beta_{21}+\beta_{22}\beta_{23}^3 & \beta_{21}+\beta_{22}\beta_{23}^4 & \beta_{21}+\beta_{22}\beta_{23}^5 \end{bmatrix},$$

$t_j = j,\ j = 1, 2, \ldots, 6;$

and

$$\Xi = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{bmatrix}. \qquad (4.3.6)$$

If, however, from previous experience, it is claimed that β_{i1}, β_{i3} are the same for both groups, then the number of parameters is reduced and need not be put in formal matrix form. Hence, in general, it is advisable to assemble parameters in a vector form. For illustration, if in the above example, $\beta_{11} = \beta_{21}$, $\beta_{13} = \beta_{23}$, then we have a new vector of parameters

$$\Xi = \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{22} \\ \beta_{13} \end{bmatrix},$$

keeping all the other quantities unchanged.

Some examples similar to the above are considered by Allen (1967). In his discussion, Ξ is always a vector of parameters i.e., $\Xi_r^T = (\beta_{11} \quad \beta_{12} \quad \beta_{13} \quad \beta_{21} \quad \beta_{22} \quad \beta_{23})$ obtained from (4.3.6). Since, in model (4.3.2), $F(\Xi)$ does not involve Ξ as an explicit matrix variable, we proceed to derive estimation equations for the vector variable Ξ_r . We observe that for any non-linear model (4.3.2), a suitable Ξ_r exists.

In this section, our objective is to present some matrix differentiation results which are useful for parameter estimation in non-linear multivariate models. Theorem 3.4.3 of Chapter III is useful in the present development.

We need the following result from matrix theory: For matrices A, B such that the traces exist,

$$\mathrm{tr}AB = \mathrm{tr}A_r({B^T}_r)^T \tag{4.3.7}$$

$$= \mathrm{tr}({A^T}_r){B_r}^T \tag{4.3.8}$$

which can be easily verified.

Now we prove a result concerning differentiable scalar functions involving $F(\Xi)$, whose elements are non-linear functions of the $\beta_{\gamma\delta}$.

<u>Theorem 4.3.1</u> If, for a differentiable scalar function $f(\Xi)$ and a differentiable matrix function $F(\Xi)$,

$$df(\Xi) = \mathrm{tr}[A(dF(\Xi))B+C(dF(\Xi))^T D], \tag{4.3.9}$$

then

$$\frac{\partial f(\Xi)}{\partial \Xi_r} = \frac{\partial (F(\Xi))_r}{\partial \Xi_r} (A^T B^T+DC)_r \ . \tag{4.3.10}$$

<u>Proof</u>: We notice that

$$d(F(\Xi))_r = \left\{\frac{\partial (F(\Xi))_r}{\partial \Xi_r}\right\}^T d\Xi_r \ . \tag{4.3.11}$$

Now

$$df(\Xi) = tr[A^TB^T(dF(\Xi))^T+DC(dF(\Xi))^T],$$

using the properties of trace functions,

$$= tr[(A^TB^T+DC)_r(dF(\Xi))_r^T], \text{ using } (4.3.7),$$

$$= tr\left[(A^TB^T+DC)_r(d\Xi_r)^T \frac{\partial(F(\Xi))_r}{\partial\Xi_r}\right], \text{ using } (4.3.11),$$

which implies

$$\frac{\partial f(\Xi)}{\partial\Xi_r} = \frac{\partial(F(\Xi))_r}{\partial\Xi_r}(A^TB^T+DC), \text{ using Theorem 3.4.3.}$$

A satisfactory estimate of Ξ is available from any one of the following procedures.

4.3.1 Least Squares Estimation

Under this procedure, we minimize

$$f(\Xi) = tr\Sigma^{-1}(X-AF(\Xi))^TV^{-1}(X-AF(\Xi)) \qquad (4.3.1.1)$$

with respect to the vector Ξ_r of parameters. On differentiating $f(\Xi)$, we obtain

$$df(\Xi) = tr\left[-\Sigma^{-1}(dF(\Xi))^TA^TV^{-1}X-\Sigma^{-1}X^TV^{-1}AdF(\Xi) + \Sigma^{-1}(dF(\Xi))^TA^TV^{-1}AF(\Xi)+\Sigma^{-1}(F(\Xi))^TA^TV^{-1}AdF(\Xi)\right].$$

On application of Theorem 4.3.1, and setting $\frac{\partial f(\Xi)}{\partial\Xi_r} = 0$, we obtain the following normal equations for the non-linear model

$$\frac{\partial(F(\Xi))_r}{\partial\Xi_r} A^TV^{-1}[X-AF(\Xi)] = 0 \qquad (4.3.1.2)$$

Any solutions $\hat{\Xi}_r$ of (4.3.1.2) provide a good initial estimate of the Ξ_r .

4.3.2 Maximum Likelihood Estimation

Assuming that U is normally distributed, the logarithm of the likelihood function is given by

$$\log L(\Xi,V,\Sigma|X) = \text{constant}-\frac{n}{2}\log|\Sigma|-\frac{p}{2}\log|V|$$

$$-\frac{1}{2}\text{tr}\Sigma^{-1}(X-AF(\Xi))^T V^{-1}(X-AF(\Xi)). \quad (4.3.2.1)$$

Proceeding as in Section 4.3.1, an approximate maximum likelihood estimate of Ξ is obtained from the likelihood equation given by (4.3.1.2). Exactly as in the linear case, if V is known, and if there exists a maximum likelihood estimate $\hat{\Xi}$ of Ξ, then from Lemma 4.2.2.1

$$\hat{\Sigma} = \frac{1}{n}(X-AF(\hat{\Xi}))^T V^{-1}(X-AF(\hat{\Xi})) \quad (4.3.2.2)$$

is a maximum likelihood estimate of Σ. If, in particular, $V = I$, expression (4.3.2.2) is similar to that given in Allen (1967, Theorem 2.4.1, (2)), with our $\hat{\Sigma}$ replaced by his $\hat{V}$. We note that $\hat{\Sigma}$ for a linear model may be obtained by substituting $F(\hat{\Xi}) = \hat{\Xi}P$ in (4.3.2.2). Allen (1967) provides a detailed proof for showing that $\hat{\Xi}$ and $\hat{\Sigma}$ are maximum likelihood estimates of Ξ and Σ, if $\hat{\Xi}$ is any value which minimizes

$$\log|(X-AF(\Xi))^T(X-AF(\Xi)|.$$

Allen (1967) obtained an initial estimate $\hat{\Xi}_r$ using a different technique. He used this $\hat{\Xi}_r$ in minimizing the (1) logarithm of generalized sample variance and (2) a trace function similar to (4.3.1.1), with $V = I$, by the method of steepest descent and a weighted linearization procedure.

4.4 Some Constrained Extremum Problems of Matrix Functions

Dwyer (1958) discussed the minimum variance unbiased estimation of a vector of linear parametric functions for the model

$$\underline{x} = A\underline{\beta}+\underline{u} \ , \tag{4.4.1}$$

where $\underline{x}$:n×1, $\underline{\beta}$:g×1 are appropriate column vectors, representing a linear regression. He applied the theory of symbolic matrix derivatives [Dwyer & MacPhail (1948)] to extend Aitken's Generalized Gauss-Markov Least Squares Theorem to a vector of linear parametric functions. Johnston (1963) obtained the minimum variance unbiased estimator of $\underline{\beta}$ using scalar differentiation. Results by Rao (1965, pp. 189-191) and Johnston (1963, pp. 180-183) are particular cases of those of Dwyer (1958, pp. 107-111). Neudecker (1967) gave a simplified proof of Johnston's result, using matrix differentiation. Tracy & Dwyer (1969, pp. 1585-1586) considered the extrema of quadratic forms under a linear restriction, using vector differentiation. All the above extremum problems were solved using Lagrangian multipliers.

In this section we extend the use of matrix differentiation theorems to discuss extreme values of matrix functions under linear restrictions on the matrix variables under a more general set-up. Results of Dwyer (1958) and Johnston (1963, pp. 182-183) are obtained as special cases of our results.

We consider the general model of Section 4.2 described by (4.2.1), (4.2.2) and (4.2.3).

4.4.1 Minimum Variance Unbiased Estimation of Ξ

Let

$$\Xi^* = BXC \tag{4.4.1.1}$$

be a linear estimator of the parameter matrix Ξ, and $B{:}g\times n$, $C{:}p\times h$ be matrices to be determined. Suppose Ξ^* is an unbiased estimator of Ξ, then

$$E(\Xi^*) = BA\Xi PC$$
$$= \Xi$$

if and only if

$$BA = I, \qquad PC = I \quad . \tag{4.4.1.2}$$

Since

$$\Xi^* = \Xi + BUC$$

we have, using Theorem 2.6.3.1,

$$\Xi^*_r = \Xi_r + (B\otimes C^T)U_r \quad . \tag{4.4.1.3}$$

The sum of the sampling variances of the estimates of Ξ^* is given by

$$E[(\Xi^*_r - \Xi_r)^T(\Xi^*_r - \Xi_r)] = E[U_r^T(B^T\otimes C)(B\otimes C^T)U_r]$$

$$= tr[(B^TB\otimes CC^T)(V\otimes\Sigma)] \quad . \tag{4.4.1.4}$$

From (4.4.1.2), we obtain

$$BA\otimes C^TP^T = (B\otimes C^T)(A\otimes P^T) = I \quad . \tag{4.4.1.5}$$

Hence for Ξ^* to be a minimum variance unbiased estimator of Ξ, we need to find $B\otimes C^T$ such that

$$f(B\otimes C^T) = tr[(B^TB\otimes CC^T)(V\otimes\Sigma)] - 2tr[(BA\otimes C^TP^T - I)\Lambda] \tag{4.4.1.6}$$

is a minimum; Λ:gh×gh being a matrix of Lagrangian multipliers.

On differentiating $f(B\otimes C^T)$ with respect to $(B\otimes C^T)$, we obtain

$$df(B\otimes C^T) = tr[\{d(B^T\otimes C)\}(B\otimes C^T)(V\otimes\Sigma)+(B^T\otimes C)\{d(B\otimes C^T)\}(V\otimes\Sigma)]$$

$$-2tr[\{d(B\otimes C^T)\}(A\otimes P^T)\Lambda] \ . \qquad (4.4.1.7)$$

Using Theorem 3.4.3 and setting $\frac{\partial f(B\otimes C^T)}{\partial(B\otimes C^T)} = 0$, we have

$$(B\otimes C^T)(V\otimes\Sigma) = \Lambda^T(A^T\otimes P) \qquad (4.4.1.8)$$

so that for any solution $(\hat{B}\otimes\hat{C}^T)$ of (4.4.1.8), $f(B\otimes C^T)$ has an extreme value.

Further differentiation of (4.4.1.7) leads to

$$d^2f(B\otimes C^T) = 2tr[\{d(B^T\otimes C)\}\{d(B\otimes C^T)\}(V\otimes\Sigma)] \ . \qquad (4.4.1.9)$$

From (4.4.1.9), on using Corollary 3.8.2,

$$\nabla^2f(B\otimes C^T) = 2I\otimes(V\otimes\Sigma) \ , \qquad (4.4.1.10)$$

which is positive definite (cf. A.1.35) and hence $f(\hat{B}\otimes\hat{C}^T)$ is a minimum, from Section 3.9.

Now we show that the minimum variance unbiased estimator Ξ^* is equal to the least-squares estimator $\hat{\Xi}$ of the matrix parameter Ξ.

On post-multiplying (4.4.1.8) by $(V^{-1}\otimes\Sigma^{-1})(A\otimes P^T)$ and using (4.4.1.5), we get

$$\Lambda^T(A^T\otimes P)(V^{-1}\otimes\Sigma^{-1})(A\otimes P^T) = I$$

i.e. $\Lambda = \{(A^T \otimes P)(V^{-1} \otimes \Sigma^{-1})(A \otimes P^T)\}^{-1}$ (4.4.1.11)

if the inverse of the right hand side exists.

Substituting (4.4.1.11) in (4.4.1.8), we get

$$B \otimes C^T = \{(A^T \otimes P)(V^{-1} \otimes \Sigma^{-1})(A \otimes P^T)\}^{-1}(A^T \otimes P)(V \otimes \Sigma)^{-1}. \quad (4.4.1.12)$$

Now, expression (4.4.1.1) yields

$$\Xi^*_r = (B \otimes C^T)X_r$$

$$= \{(A^T \otimes P)(V^{-1} \otimes \Sigma^{-1})(A \otimes P^T)\}^{-1}(A^T \otimes P)(V \otimes \Sigma)^{-1}X_r \quad (4.4.1.13)$$

which implies

$$(A^T V^{-1} A \otimes P \Sigma^{-1} P^T)\Xi^*_r = (A^T V^{-1} \otimes P \Sigma^{-1})X_r \ . \quad (4.4.1.14)$$

From (4.4.1.14), we have

$$A^T V^{-1} A \Xi^* P \Sigma^{-1} P^T = A^T V^{-1} X \Sigma^{-1} P^T \ , \quad (4.4.1.15)$$

which is the same as the normal equation (4.2.1.4) for the least squares estimator $\hat{\Xi}$.

If, in particular, $h = p = 1$, $P = 1$ in the general model (4.2.1), we get the model (4.4.1) considered by Johnston (1963, p. 179). Consequently, $C = 1$, $\Sigma = 1$, $\Lambda : g \times g$, the matrix of Lagrangian multipliers, would transform the results of this section to those of Johnston (1963, pp. 182-183) and of Neudecker (1967, pp. 106-107). The extremum problem considered by Tracy & Dwyer (1969, p. 1586) involved a vector of Lagrangian multipliers and may be treated as a special case of Johnston's result by using the substitution

$$(A, V, X, I, \Lambda) \rightarrow (\underline{x}^T, A, B, \underline{u}^T, \underline{\lambda}) \ .$$

4.4.2 Minimum Variance Unbiased Estimation of a Linear Matrix Function of Ξ

A simpler proof of the constrained minimization problem considered by Dwyer (1958) may be given using Theorem 3.4.3 and Section 3.8. The aim of this subsection is to extend Dwyer's result to obtain a restricted minimum variance unbiased estimator of a linear parametric function for a general multivariate model.

Consider a linear matrix function

$$\Phi = L\Xi M \qquad (4.4.2.1)$$

where $L{:}n_1 \times g$, $M{:}h \times n_2$ are known matrices.

Define

$$\Phi^* = GXH = GA\Xi PH + GUH \qquad (4.4.2.2)$$

where $G{:}n_1 \times n$, $H{:}p \times n_2$ are matrices to be determined such that Φ^* is a minimum variance unbiased estimator of Φ together with a specific restriction to be imposed below.

Now

$$E(\Phi^*) = \Phi \qquad (4.4.2.3)$$

implies, since $E(U) = 0$,

$$GA\Xi PH = L\Xi M \qquad (4.4.2.4)$$

if and only if

$$GA = L \quad \text{and} \quad PH = M. \qquad (4.4.2.5)$$

Suppose further that an additional set of restrictions is given by

$$\Psi = R\Xi Q = 0 \qquad (4.4.2.6)$$

where $R{:}n_3 \times g$, $Q{:}h \times n_4$, $gh > n_3 n_4$. These restrictions may be independent of unbiasedness and the minimum variance property of Φ. Pre-multiplying and post-multiplying (4.4.2.6) by $C:n_1\times n_3$, $D:n_4\times n_2$, matrices to be determined, we obtain

$$(CR\otimes D^TQ^T)\Xi_r = 0 \quad \text{and} \quad (C\otimes D^T)\Psi_r = 0. \tag{4.4.2.7}$$

From (4.4.2.5), we have

$$GA\otimes H^TP^T = L\otimes M^T$$

which implies

$$(G\otimes H^T)(A\otimes P^T)-L\otimes M^T = 0. \tag{4.4.2.8}$$

From (4.4.2.2), under condition (4.4.2.5), we get

$$\Phi^*{}_r-\Phi_r = (G\otimes H^T)U_r \ . \tag{4.4.2.9}$$

Hence the sum of the sampling variances of elements of Φ^* is given by

$$E[(\Phi^*{}_r-\Phi_r)^T(\Phi^*{}_r-\Phi_r)] = E[U_r{}^T(G^T\otimes H)(G\otimes H^T)U_r]$$

$$= tr[(G^T\otimes H)(G\otimes H^T)(V\otimes\Sigma)] \ . \tag{4.4.2.10}$$

Now we need to minimize (4.4.2.10) subject to the following conditions:

$$(G\otimes H^T)(A\otimes P^T)-L\otimes M^T+(C\otimes D^T)(R\otimes Q^T) = 0 \tag{4.4.2.11}$$

and

$$(C\otimes D^T)\Psi_r = 0 \ . \tag{4.4.2.12}$$

We note that the matrices C and D preserve the conformability of (4.4.2.11).

If $\Lambda:gh\times n_1n_2$ and $\Gamma:1\times n_1n_2$ are matrices of Lagrangian multipliers, then we minimize

$$f(G\otimes H^T,\ C\otimes D^T) = tr[(G^T\otimes H)(G\otimes H^T)(V\otimes\Sigma)]$$

$$-2tr[(G\otimes H^T)(A\otimes P^T)-L\otimes M^T+(C\otimes D^T)(R\otimes Q^T)]\Lambda$$

$$+2tr(C\otimes D^T)\Psi_r\Gamma \qquad (4.4.2.13)$$

with respect to $G\otimes H^T$ and $C\otimes D^T$.

On differentiating with respect to $G\otimes H^T$ and $C\otimes D^T$, and setting $\frac{\partial f}{\partial(G\otimes H^T)} = 0$, $\frac{\partial f}{\partial(C\otimes D^T)} = 0$, we obtain the following estimating equations for an extreme value

$$(G\otimes H^T)(V\otimes\Sigma) = \Lambda^T(A^T\otimes P) \qquad (4.4.2.14)$$

$$\Lambda^T(R^T\otimes Q) = \Gamma^T\Psi_r{}^T$$

$$= 0,\ \text{using } (4.4.2.6) \qquad (4.4.2.15)$$

It is interesting to compare (4.4.2.14) and (4.4.2.15) with expressions (3.11) and (3.12), respectively, in Dwyer (1958) paper. Since $V\otimes\Sigma$ is positive definite, we have, from expression (4.4.2.14),

$$G\otimes H^T = \Lambda^T(A^T\otimes P)(V^{-1}\otimes\Sigma^{-1}) \ . \qquad (4.4.2.16)$$

Post-multiplying (4.4.2.11) by $\Xi^*{}_r$ yields

$$\Phi^* = L\Xi^*M = \Lambda^T A^T V^{-1} A\Xi^* P\Sigma^{-1}P^T \qquad (4.4.2.17)$$

using (4.4.2.7), (4.4.2.16) and Theorem 2.6.3.1. Again, from expression (4.4.2.2),

$$\Phi^*{}_r = (G\otimes H^T)X_r$$

$$= \Lambda^T(A^T \otimes P)(V^{-1} \otimes \Sigma^{-1})X_r \text{ , using (4.4.2.16)}$$

Hence

$$\Phi^* = \Lambda^T A^T V^{-1} X \Sigma^{-1} P^T \tag{4.4.2.18}$$

Equating (4.4.2.17) and (4.4.2.18), we get

$$\Lambda^T A^T V^{-1} X \Sigma^{-1} P^T = \Lambda^T A^T V^{-1} A \Xi^* P \Sigma^{-1} P^T , \tag{4.4.2.19}$$

which is the same as the normal equation for least squares estimation except for a pre-factor Λ^T. Dwyer (1958, (3.17)) provides an expression for the special case which is obtained from (4.4.2.19) by substituting $X = \underline{x}:n\times 1$, $P = 1$, $\Sigma = 1$ and $\Xi^* = \underline{\theta}^*:g\times 1$.

Dwyer (1958) has not considered the second order condition for the vector form of the linear regression model. Here we establish that the extreme value of $f(G\otimes H^T,\ C\otimes D^T)$ given in (4.4.2.13) is, in fact, a minimum. Applying Corollary 3.8.2,

$$\nabla^2_{W_r, W_r} f(W,\ Z) = 2I\otimes(V\otimes\Sigma) \tag{4.4.2.20}$$

$$\nabla^2_{W_r, Z_r} f(W,\ Z) = 0 \tag{4.4.2.21}$$

$$\nabla^2_{Z_r, Z_r} f(W,\ Z) = 0, \tag{4.4.2.22}$$

where $W = G\otimes H^T$, $Z = C\otimes D^T$. Since the Hessian matrix with respect to $G\otimes H^T$ is given by $2I\otimes(V\otimes\Sigma)$, which is positive definite, and hence $f(G\otimes H^T,\ C\otimes D^T)$ is minimum at $G\otimes H^T = \Lambda^T(A^T\otimes P)(V^{-1}\otimes\Sigma^{-1})$. With the use of the present matrix differentiation theorems, one can discuss constrained extremum problems for much more complicated matrix functions.

4.5 Estimation and the Asymptotic Covariance Matrix in the Structural Econometric Model

In this section we give some applications of partitioned matrix differentiation discussed in Section 3.10. We consider a general linear structural model of simultaneous equations

$$BY+\Gamma Z = U \tag{4.5.1}$$

where $Y:p\times n$ is the matrix of observations on p jointly dependent variables; $Z:q\times n$ is the matrix of observations on the independent variables; $B:p\times p$, $\Gamma:p\times q$ are matrices of parameters, and $U:p\times n$ is the matrix of random errors. We assume that

$$E(U) = 0, \quad E(\tfrac{1}{n}UU^T) = \Sigma:p\times p \tag{4.5.2}$$

and that the matrices B and Σ are non-singular. Full information maximum likelihood (FIML) estimators of B and Γ for the above model were obtained by Fisk (1967, Chapter 4). He obtained likelihood equations by differentiating the log-likelihood functions with respect to B and Γ separately. Since the likelihood equations are non-linear in the unknown parameters, he suggested an iterative procedure for obtaining the maximum likelihood estimators of these parameters. Neudecker (1969b) discussed the above problem by using non-partitioned matrix differentiation theorems, but his expressions are very complicated and have an error. Rothenberg & Leenders (1964) have also considered FIML estimators in a model similar to the above model (4.5.1).

In a recent paper, Tracy & Singh (1971b) simplified some of the expressions in the paper of Neudecker (1969b, p. 962), using partitioned matrix differentiation. The aim of this section is to consider the structural parameter estimation of (4.5.1) without assuming that U is normally distributed. A new procedure for obtaining the asymptotic covariance matrix in a simultaneous equation model is also suggested.

4.5.1 Full Information - Least Generalized Residual Variance (FI/LGRV) Estimators

We can express model (4.5.1) as

$$AW = U \,, \tag{4.5.1.1}$$

which is a matrix function involving the partitioned matrices

$A = (B \quad \Gamma)$, $W = \begin{bmatrix} Y \\ Z \end{bmatrix}$. Since Y is a dependent variable matrix,

we pre-multiply (4.5.1) by B^{-1} to get the reduced form

$$\begin{aligned} Y &= -B^{-1}\Gamma Z + B^{-1}U \\ &= \Pi Z + V \end{aligned} \tag{4.5.1.2}$$

where $\Pi = -B^{-1}\Gamma$, $V = B^{-1}U$. Since expression (4.5.1.2) is a linear regression of Y on Z, the full information - least generalized residual variance method leads to minimizing the function $\frac{1}{2}\log|VV^T|$ with respect to A. Now

$$\begin{aligned} f(A) = \tfrac{1}{2}\log|VV^T| &= \tfrac{1}{2}\log|B^{-1}AWW^TA^TB^{-T}| \\ &= -\log|B| + \tfrac{1}{2}\log|AMA^T| \,, \end{aligned} \tag{4.5.1.3}$$

where $M = \begin{bmatrix} YY^T & YZ^T \\ ZY^T & ZZ^T \end{bmatrix}$. Goldberger (1964, p. 353) mentions

difficulty in proceeding with the minimization of (4.5.1.3).

He, however, provides references for computational procedure.

We discuss the minimization of (4.5.1.3) using our results for partitioned matrix differentiation. Differentiating f(A) with respect to the partitioned matrix A, we have

$$df(A) = -\mathrm{tr}\, B^{-T}dB^{T}+\tfrac{1}{2}\mathrm{tr}[(AMA^{T})^{-1}\{(dA)MA^{T}+AM(dA^{T})\}]$$

$$= -\mathrm{tr}[B^{-T} \vdots \; O]dA^{T}+\mathrm{tr}(AMA^{T})^{-1}(dA)MA^{T} \ , \qquad (4.5.1.4)$$

using a trace property.

Applying Theorem 3.4.3, the first order matrix derivative is

$$\frac{\partial f(A)}{\partial A} = -[B^{-T} \vdots \; O]+(AMA^{T})^{-1}AM \ . \qquad (4.5.1.5)$$

Setting $\frac{\partial f(A)}{\partial A} = O$, we obtain A from

$$\begin{bmatrix} \hat{B}^{-1} \\ \cdots \\ O \end{bmatrix} = M\hat{A}^{T}(\hat{A}M\hat{A}^{T})^{-1} \ . \qquad (4.5.1.6)$$

Since (4.5.1.6) is non-linear in the unknown parameter A, we need to evaluate the Hessian matrix which is required in the Newton-Raphson and Fletcher-Powell methods for estimating A.

Let

$$F(A) = \frac{\partial f(A)}{\partial A} \ ; \text{ then } \ \nabla^{2}_{A_R,A_R} f(A) = \frac{\partial F_R}{\partial A_R} \ , \qquad (4.5.1.7)$$

where $[\ ]_R$ denotes the column vector representation as in Definition 2.10.1, e.g. $A_R = \begin{bmatrix} B_r \\ \cdots \\ \Gamma_r \end{bmatrix}$.

On differentiating F(A), we get

$$dF(A) = [B^{-T} \vdots \; O](dA^{T})[B^{-T} \vdots \; O]-(AMA^{T})^{-1}(dA)MA^{T}(AMA^{T})^{-1}AM$$

$$-(AMA^T)^{-1}AM(dA^T)(AMA^T)^{-1}AM+(AMA^T)^{-1}(dA)M \ . \tag{4.5.1.8}$$

Since expression (4.5.1.8) is a partitioned matrix function, we can apply the results of partitioned matrix differentiation very easily. Applying Theorems 2.10.2 and 3.10.1 to (4.5.1.8), we obtain the following expression involving the ⓟ product

$$dF_R = \left\{[B^{-T} \vdots \ 0] \ ⓟ \begin{bmatrix} B^{-1} \\ \cdots \\ 0 \end{bmatrix}\right\}(I_{\{2\}})dA_R$$

$$-\{(AMA^T)^{-1} ⓟ \ MA^T(AMA^T)^{-1}AM\}dA_R$$

$$-\{(AMA^T)^{-1}AM ⓟ \ MA^T(AMA^T)^{-1}\}(I_{\{2\}})dA_R$$

$$+\{(AMA^T)^{-1} ⓟ \ M\}dA_R \ . \tag{4.5.1.9}$$

For the product ⓟ , we have

$$[B^{-T} \vdots \ 0] ⓟ \begin{bmatrix} B^{-1} \\ \cdots \\ 0 \end{bmatrix} = \begin{bmatrix} B^{-T}\otimes B^{-1} & \vdots & 0 \\ \cdots\cdots & \vdots & \cdots \\ 0 & \vdots & 0 \end{bmatrix} .$$

Since in this problem $(I_{\{2\}})^T = {}_{\{1\}}I = I$, an application of (3.10.1) yields

$$\nabla^2_{A_R,A_R} f(A)\Big|_{A=\hat{A}} = \begin{bmatrix} \hat{B}^{-1}\otimes\hat{B}^{-T} & \vdots & 0 \\ \cdots\cdots & \vdots & \cdots \\ 0 & \vdots & 0 \end{bmatrix} - \{(\hat{A}M\hat{A}^T)^{-1} ⓟ \ M\hat{A}^T(\hat{A}M\hat{A}^T)^{-1}\hat{A}M\}$$

$$-M\hat{A}^T(\hat{A}M\hat{A}^T)^{-1} ⓟ \ (\hat{A}M\hat{A}^T)^{-1}\hat{A}M+\{(\hat{A}M\hat{A}^T)^{-1} ⓟ \ M\} \ , \tag{4.5.1.10}$$

which is the Hessian matrix for any solution $\hat{A}$ of A.

It is interesting to note that expressions (4.5.1.6) and (4.5.1.10) are exactly the same as the corresponding results obtained on assuming that the matrix U is normally distributed [see Tracy & Singh (1971b)]. Expression (4.5.1.10) may be compared to that of Fisk (1967, p. 31, (4.12)). In Neudecker (1969b, p. 962), the first term on the right hand side of the expression for $\nabla^2_{\text{vec } A} L$ should be

$$\begin{bmatrix} [B^{-T} \vdots 0] \otimes \begin{bmatrix} B^{-1} \\ \cdots \\ 0 \end{bmatrix}_{1.} \\ \vdots \\ [B^{-T} \vdots 0] \otimes \begin{bmatrix} B^{-1} \\ \cdots \\ 0 \end{bmatrix}_{(p+q).} \end{bmatrix} = \left[\begin{array}{c:c} B^{-T} \otimes (B^{-1})_{1.} & \\ \vdots & 0 \\ B^{-T} \otimes (B^{-1})_{p.} & \\ \hdashline 0 & 0 \end{array}\right] \qquad (4.5.1.11)$$

instead of

$$\begin{bmatrix} [B^{-T} \vdots 0] \otimes (B^{-1})_{1.} \\ \vdots \\ [B^{-T} \vdots 0] \otimes (B^{-1})_{p.} \end{bmatrix} . \qquad (4.5.1.12)$$

The approach followed by the above authors for estimating the unknown parameter matrix A is known as the <u>full information maximum likelihood</u> method.

4.5.2 Asymptotic Covariance Matrix

Sometimes we come across a non-linear matrix function $Y = F(X)$ of a matrix variable X with known mean and covariance matrices. Then the expressions for asymptotic mean and covariance matrices of Y are obtained in terms of the mean and covariance matrices of X. These asymptotic expressions extend a result of Goldberger (1964, p. 125) to matrix functions. Below we present these asymptotic expressions for non-partitioned matrix functions. Corresponding expressions for the partitioned case are obtained from the non-partitioned one.

Theorem 4.5.2.1 Consider a non-linear differentiable matrix function

$$Y = F(X) \ , \quad Y{:}p\times q, \ X{:}m\times n \tag{4.5.2.1}$$

with

$$E(X) = \Xi, \quad Var(X_r) = \Sigma{:}mn\times mn. \tag{4.5.2.2}$$

Then

$$E(Y) = F(\Xi) \tag{4.5.2.3}$$

and

$$Var(Y_r) = \left(\frac{\partial Y_r}{\partial X_r}\right)^T \Sigma \left(\frac{\partial Y_r}{\partial X_r}\right) \ {:}pq\times pq \ . \tag{4.5.2.4}$$

The proof of this theorem follows immediately from Goldberger (1964, pp. 123-125).

The results of Theorem 4.5.2.1 are extended to partitioned matrix functions on replacing Y_r, X_r by Y_R, X_R respectively. For example, an expression for the asymptotic covariance matrix for the partitioned situation is given by

$$\mathrm{Var}(Y_R) = \left(\frac{\partial Y_R}{\partial X_R}\right)^T \Sigma \left(\frac{\partial Y_R}{\partial X_R}\right) . \qquad (4.5.2.5)$$

Clearly, in the reduced form (4.5.1.2) of the structural model (4.5.1), $\hat{\Pi} = -\hat{B}^{-1}\hat{\Gamma}$ is a non-linear partitioned matrix function of $\hat{A} = (\hat{B} \vdots \hat{\Gamma})$. We apply expression (4.5.2.5) to obtain the asymptotic covariance matrix $\mathrm{Var}(\hat{\Pi}_r)$ in terms of $\mathrm{Var}(\hat{A}_R)$.

Now suppose

$$\mathrm{Var}(\hat{A}_R) = \begin{bmatrix} \mathrm{Var}(\hat{B}_r) & \vdots & \mathrm{Cov}(\hat{B}_r, \hat{\Gamma}_r) \\ \cdots\cdots & \vdots & \cdots\cdots \\ \mathrm{Cov}(\hat{\Gamma}_r, \hat{B}_r) & \vdots & \mathrm{Var}(\hat{\Gamma}_r) \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\Delta}_{11} & \vdots & \hat{\Delta}_{12} \\ \cdots & \vdots & \cdots \\ \hat{\Delta}_{21} & \vdots & \hat{\Delta}_{22} \end{bmatrix} \qquad (4.5.2.6)$$

$$= \hat{\Delta} \qquad (4.5.2.7)$$

is given. Since

$$\Pi = -B^{-1}\Gamma ,$$

we obtain, on differentiation,

$$d\Pi = B^{-1}(dB)B^{-1}\Gamma - B^{-1}(d\Gamma)$$

$$= -B^{-1}[dB \vdots d\Gamma] \begin{bmatrix} \Pi \\ \cdots \\ I \end{bmatrix} . \qquad (4.5.2.8)$$

Since Π is a non-partitioned matrix, $\Pi_R = \Pi_r$. Using Theorem 2.10.2, we have

$$d\Pi_r = -\{B^{-1} \circledast [\Pi^T \vdots I]\} dA_R , \qquad (4.5.2.9)$$

where

$$dA_R = \begin{bmatrix} dB_r \\ \cdots \\ d\Gamma_r \end{bmatrix} . \tag{4.5.2.10}$$

Now, by applying Theorem 3.10.1 to (4.5.2.9), we identify the matrix of partial derivatives as

$$\frac{\partial \Pi_r}{\partial A_R} = \begin{bmatrix} \frac{\partial \Pi_r}{\partial B_r} \\ \cdots \\ \frac{\partial \Pi_r}{\partial \Gamma_r} \end{bmatrix} = -B^{-T} \textcircled{\pi} \begin{bmatrix} \Pi \\ \cdots \\ I \end{bmatrix} . \tag{4.5.2.11}$$

Hence

$$\left.\frac{\partial \Pi_r}{\partial A_R}\right|_{A=\hat{A}} = -\begin{bmatrix} \hat{B}^{-T} \otimes \hat{\Pi} \\ \cdots\cdot \\ \hat{B}^{-T} \otimes I \end{bmatrix} . \tag{4.5.2.12}$$

The asymptotic covariance matrix of $\hat{\Pi}_r$ is

$$\mathrm{Var}(\hat{\Pi}_r) = [\hat{B}^{-1} \otimes \hat{\Pi}^T \vdots \hat{B}^{-1} \otimes I]\hat{\Delta} \begin{bmatrix} \hat{B}^{-T} \otimes \hat{\Pi} \\ \cdots\cdot \\ \hat{B}^{-T} \otimes I \end{bmatrix} , \tag{4.5.2.13}$$

using expression (4.5.2.5). The process followed by Goldberger, Nagar & Odeh (1961, pp. 560-561); see also Goldberger (1964, pp. 370-371), for obtaining $\mathrm{Var}(\hat{\Pi}_r)$ is based on elementwise differentiation of Π with respect to the elements of A. Their proof was simplified to some extent by Neudecker (1968, p. 74) where he identifies $\frac{\partial \Pi_r}{\partial B_r}$ and $\frac{\partial \Pi_r}{\partial \Gamma_r}$ separately and then assembles them to obtain an expression for $\frac{\partial \Pi_r}{\partial A_r}$. The proof followed in this section is much easier than those followed by above authors. Our procedure, besides possessing simplicity, has certain additional advantages. For the sake of illustration, suppose that the elements of $\hat{B}$ are uncorrelated with those of $\hat{\Gamma}$. Then $\hat{\Delta}_{12} = \hat{\Delta}_{21} = 0$ and an explicit expression for $\mathrm{Var}(\hat{\Pi}_r)$ is given by

$$\mathrm{Var}(\hat{\Pi}_r) = [\hat{B}^{-1}\otimes\hat{\Pi}^T \vdots \hat{B}^{-1}\otimes I]\begin{bmatrix}\hat{\Delta}_{11} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & \hat{\Delta}_{22}\end{bmatrix}\begin{bmatrix}\hat{B}^{-T}\otimes\hat{\Pi} \\ \cdots \\ \hat{B}^{-T}\otimes I\end{bmatrix}$$

$$= (\hat{B}^{-1}\otimes\hat{\Pi}^T)\hat{\Delta}_{11}(\hat{B}^{-T}\otimes\hat{\Pi})+(\hat{B}^{-1}\otimes I)\hat{\Delta}_{22}(\hat{B}^{-T}\otimes I). \quad (4.5.2.14)$$

Following an approach given by Goldberger (1964, p. 370) or by Neudecker (1968), the covariance matrix of A is not block diagonal under the above assumption. Hence in this case an explicit expression for the covariance matrix of $\hat{\Pi}$ in terms of the covariance matrices of $\hat{B}_r$ and $\hat{\Gamma}_r$ is much more complicated.

4.6 Estimation in the Analysis of Covariance Structures

The purpose of this section is to discuss estimation problems in the analysis of covariance structures with the help of the matrix derivative formulae developed in Sections 3.4 and 3.8. First we give a general description of the covariance structural model in the paper of Jöreskog (1970a). We then obtain:

(1) Three sets of estimation equations for the unknown structural parameter matrices using the Minres function, the Likelihood function and a generalized Howe's function.

(2) The Hessian matrices of the Minres function/Likelihood function which can be tested for positive/negative definiteness at the solutions of the above estimation equations. These are also required in the Fletcher & Powell (1963) method and the Newton-Raphson iterative procedure for finding improved estimates of the unknown structural parameter matrices for the above functions.

(3) Estimates of the non-random latent vectors of the general covariance structural model.

Estimation problems in factor analysis have been considered by several authors. Anderson & Rubin (1956) and Lawley & Maxwell (1963) have given a detailed treatment of the estimation of parameter matrices and of factor scores in a factor analysis model. Some improvements in the estimation of parameters were made by Bock & Bargmann (1966), Jöreskog (1966, 1967, 1969), Jöreskog & Lawley (1968), and Lawley (1967), all based on the method of maximum likelihood. An alternative estimation procedure was suggested by Howe [see Morrison (1967, pp. 286-289)]. The above authors obtained the estimating equations by arranging the typical elements of the first order partial derivatives in matrix form.

The factor analysis model considered by the above authors was generalized by Jöreskog (1970a) by giving a general parametric structure to the mean vector and covariance matrix of the vector of response variables. He obtained the first order partial derivatives of the log-likelihood function with respect to parameter matrices. However, he does not mention the method used for differentiation. In a recent paper, Jöreskog (1971) pointed out certain applications of the general model without any parametric structure of the mean vector.

Matrix derivatives were found very appropriate for estimating the unknown parameters of the factor analysis model by minimizing or maximizing various goodness of fit functions. To start with, likelihood equations for a model in the paper of Jöreskog (1966) have been given by Dwyer (1967) using matrix derivatives. Further attempts were made by Mulaik(1971) and Gebhardt (1971) to develop some matrix differentiation formulae to obtain estimates of the unknown parameters in factor analysis. Mulaik (1971) reported first and second order partial derivatives of both the maximum likelihood and the least squares goodness-of-fit functions of a factor analysis model in Jöreskog's (1969) paper. Also, he pointed out certain limitations of Newton-Raphson method and suggested the application of second derivatives in those situations where this method is preferable. An extension of the model in Jöreskog (1969) has been considered by Gebhardt (1971). He completely specifies some unknown parameters in advance and then applies his matrix formulae to obtain the maximum likelihood estimates of the remaining parameters by using the iterative procedure of Fletcher & Powell (1963) or by a gradient method.

Under the basic assumptions of a covariance structural model, all the goodness-of-fit functions to be minimized/maximized for estimation purposes are scalar functions of parameter matrices as arguments. Some of these are symmetric or diagonal matrices. In their approach, Jöreskog (1970a) and Mulaik (1971) assumed that $\frac{\partial f(X)}{\partial x_{ij}} = \frac{\partial f(X)}{\partial x_{ji}}$ and $\frac{\partial x_{ij}}{\partial x_{ji}} = 1$, when $i \neq j$, as scalar elements, when X is a symmetric matrix.

Anderson & Rubin (1956, pp. 140-141) obtained first order partial derivatives of the log-likelihood function with respect to the typical elements of the parameter matrices for the ordinary factor analysis model. For differentiation purposes they used the fact that $<x_{ij}> \neq <x_{ji}>$, even if $X = X^T$. This particular fact has been incorporated by Dwyer (1967, p. 608), Tracy & Dwyer (1969, pp. 1578-1582) and Gebhardt (1971, p. 157) in their general matrix differentiation formulae, which have applications in multivariate analysis.

In our approach, we postulate that for any matrix variable any two of its matrix elements are treated as two variables with no equality relations. This postulate is meaningful even if these two matrix elements have equal numerical values, since they differ in their identity because of their different positions in the matrix. Under this definition we obtain very general matrix derivative results in a simpler way, see Dwyer (1970). These results ignore any known or unknown relationships, equality or otherwise, among the (scalar) elements of a matrix. Matrix derivatives of functions of matrices with respect to their argument matrices, some of whose elements are constant e.g., diagonal matrices and/or those which are related functionally or otherwise, may be obtained from the general results. We substitute dummy matrices X^* in place of special matrices X (e.g., symmetric and diagonal matrices X) to obtain very general matrix derivative results for f(X). The asterisk notation has been used earlier (Section 4.4) in a different context.

Replacement of special matrices by dummy matrices is continued until the second order partial matrix derivatives of the goodness-of-fit functions under consideration are evaluated. After the second derivatives are obtained, any possible algebraic simplification may be carried out. These derivatives are used to obtain a unique Hessian matrix which is a very general result concerning second derivatives. Appropriate modifications in the general results may be made to obtain various special results for statistical applications.

A few special results may be obtained by using the following terminology:

(1) For a diagonal matrix $X{:}m\times m$, N_X denotes the $m\times m^2$ matrix whose respective rows are the 1 mod(m+1)-th rows of I_{m^2}, and

(2) for a diagonal matrix X, X_d denotes the column vector of diagonal matrix elements.

4.6.1 The General Covariance Structural Model

Consider the covariance structural model

$$\underline{x} = B\underline{y}+\underline{\varepsilon} \tag{4.6.1.1}$$

where $x{:}p\times 1$ is a vector of observable random variables, $B{:}p\times q$ is a matrix of unknown coefficients, $\underline{y}{:}q\times 1$ $(q<p)$ is a vector of unobserved random variables and $\underline{\varepsilon}{:}p\times 1$ is a vector of residuals. We may further represent $\underline{y}$ as

$$\underline{y} = \Lambda\underline{z}+\underline{u} \tag{4.6.1.2}$$

where $\Lambda{:}q\times k$ is a matrix of unknown coefficients, $\underline{z}{:}k\times 1$ $(k<q)$ is a vector of unobserved random variables with respect to $\underline{y}$ and $\underline{u}$ is the corresponding vector of q residuals. We assume that

(i) $E(\underline{u}) = \underline{0}$ and $E(\underline{\varepsilon}) = \underline{0}$

(ii) $Var(\underline{z}) = \Phi$, $Var(\underline{u}) = \Psi$, $Var(\underline{\varepsilon}) = \Theta$
where Ψ and Θ are diagonal matrices

(iii) $\underline{\varepsilon}$, $\underline{y}$ are uncorrelated, and $\underline{u}$, $\underline{z}$ are uncorrelated.

From these assumptions, a fundamental representation of the covariance matrix Σ of $\underline{x}$ is given by the structural form

$$\Sigma = B(\Lambda\Phi\Lambda^T+\Psi)B^T+\Theta \ . \tag{4.6.1.3}$$

Let $X{:}n\times p$ be the observation matrix whose rows are uncorrelated with the same covariance matrix Σ. Then a generalized analysis of variance model of Potthoff & Roy (1964) is given by

$$E(X) = A\Xi P \qquad (4.6.1.4)$$

where $\Xi: g\times h$ is an unknown matrix of parameters, $A: n\times g$ is a model matrix of rank $r \leqslant p \leqslant n$ and $P: h\times p$ is a known matrix of rank $h \leqslant p$. For any unknown $\Xi: g\times h$, we define $p\times p$ symmetric matrices

$$T = \frac{1}{n}(X-A\Xi P)^T(X-A\Xi P) \qquad (4.6.1.5)$$

and

$$S = \frac{1}{n-1}(X-\bar{X})^T(X-\bar{X}) \ . \qquad (4.6.1.6)$$

The statistical problem of parameter estimation in the covariance structural analysis is solved by estimating B, Λ, Φ, Ψ, Θ and Ξ such that Σ is fitted to T or S. If, on the other hand, we assume that the vectors $\underline{y}$ and $\underline{z}$ are unobserved non-random variables, then $\underline{y}$ and $\underline{z}$ enter in the model as vectors of parameters which may also be estimated.

We consider three methods of estimating the unknown parameter matrices in Σ. After these parameter matrices are estimated, a simple procedure for estimating $\underline{y}$ and $\underline{z}$ is also given.

4.6.2 The Minres Function

For a simple factor analysis model [Jöreskog (1969)], a distribution-free estimation procedure was suggested by Mulaik (1971). In this section we extend this procedure to obtain parameter estimates of the general covariance structural model by minimizing the sum of squares of the residuals such that Σ is fitted to S. This minimizing quantity may be treated as the minres function

$$f(B, \Lambda, \Phi^*, \Psi^*, \Theta^*) = tr[(\Sigma^*-S)^2]. \qquad (4.6.2.1)$$

It may be noted that minres function is not scale-free.

In particular, if in the general model $\Xi = 0$, then (4.6.2.1) reduces to the least squares function considered by Jöreskog (1971, p. 110) for which he did not provide any theoretical details. However, Jöreskog, Gruvaeus & van Thillo (1970) have written a general computer program to estimate the parameters by minimizing the above least squares function. If, in addition, $B = I$, $\Theta = 0$, then we obtain a goodness of fit criterion discussed by Mulaik (1971, p. 64). He obtained the typical elements of the first and second order derivatives of

$$f(\Lambda, \Phi^*, \Psi^*) = tr[(\Sigma^*-S)(\Sigma^*-S)^T] \qquad (4.6.2.2)$$

which may be used in the iterative procedure for improving the estimates.

We obtain estimating equations and the Hessian matrix for a model in Mulaik (1971, p. 63). We treat this special case only for the sake of simple illustration. Results for the general model may similarly be obtained by treating the <u>minres function</u> (4.6.2.1).

A certain basic formulation of the minres solution is given by Harman (1967, Chapter IX).

4.6.2.1 <u>Least Squares Estimates</u>

Our aim is to minimize a particular minres function

$$f(\Lambda, \Phi^*, \Psi^*) = tr[(\Sigma^*-S)^2] , \qquad (4.6.2.3)$$

where $\Sigma^{*} = \Lambda\Phi^{*}\Lambda^{T}+\Psi^{*}$.

Necessary conditions for a minimum are obtained by differentiating (4.6.2.3) with respect to parameter matrices Λ, Φ^{*}, and Ψ^{*}. This leads to

$$df(\Lambda, \Phi^{*}, \Psi^{*}) = 2tr(\Sigma^{*}-S)[\{(d\Lambda)\Phi^{*}\Lambda^{T}+\Lambda\Phi^{*}(d\Lambda^{T})\}+\Lambda(d\Phi^{*})\Lambda^{T} + d\Psi^{*}]. \quad (4.6.2.4)$$

Using Theorem 3.4.3, we obtain first order matrix derivatives by replacing dummy matrices by the corresponding original matrices as

$$\frac{\partial f}{\partial \Lambda} = 4(\Sigma-S)\Lambda\Phi \quad (4.6.2.5)$$

$$\frac{\partial f}{\partial \Phi} = 2\Lambda^{T}(\Sigma-S)\Lambda \quad (4.6.2.6)$$

$$\frac{\partial f}{\partial \Psi} = 2(\Sigma-S)\times\times I. \quad (4.6.2.7)$$

Least squares estimates $\hat{\Lambda}$, $\hat{\Phi}$, and $\hat{\Psi}$ are any solutions of the following equations

$$\hat{\Sigma}\hat{\Lambda}\hat{\Phi} = S\hat{\Lambda}\hat{\Phi} \quad (4.6.2.8)$$

$$\hat{\Lambda}^{T}\hat{\Sigma}\hat{\Lambda} = \hat{\Lambda}^{T}S\hat{\Lambda} \quad (4.6.2.9)$$

$$\hat{\Sigma}\times\times I = S\times\times I \quad (4.6.2.10)$$

which are obtained by setting the matrix derivatives $\frac{\partial f}{\partial \Lambda}$ etc. equal to zero matrix. Our expressions (4.6.2.5), (4.6.2.6) and (4.6.2.7) are different from (15a), (15b) and (15c) in Mulaik's paper (1971, pp. 68-69), since our approach is different. In our approach, we use the concept of matrix element of a matrix, which yields more general results.

4.6.2.2 The Hessian Matrix for the Minres Function

The approach followed by Mulaik (1971) in obtaining second derivatives of fitting functions in factor analysis is very complicated. We apply Theorem 3.8.1 and its corollaries to obtain the matrix form of the second order and the mixed partial matrix derivatives for the minres function (4.6.2.3). Since this particular minres function involves only three parameter matrices, we need to calculate only three second order and three mixed partial derivative matrices.

Differentiating (4.6.2.3) twice by treating it as a function of the parameter matrix Λ only, we have

$$
\begin{aligned}
d^2f(\Lambda) &= 2\mathrm{tr}[d(\Sigma^*-S)\{(d\Lambda)\Phi^*\Lambda^T+\Lambda\Phi^*(d\Lambda^T)\} \\
&\quad +(\Sigma^*-S)\{(d\Lambda)\Phi^*(d\Lambda^T)+(d\Lambda)\Phi^*(d\Lambda^T)\}] \\
&= 2\mathrm{tr}[\{(d\Lambda)\Phi^*\Lambda^T(d\Lambda)\Phi^*\Lambda^T+(d\Lambda)\Phi^*\Lambda^T\Lambda\Phi^*(d\Lambda^T) \\
&\quad +\Lambda\Phi^*(d\Lambda^T)(d\Lambda)\Phi^*\Lambda^T+\Lambda\Phi^*(d\Lambda^T)\Lambda\Phi^*(d\Lambda^T)\} \\
&\quad +(\Sigma^*-S)\{(d\Lambda)\Phi^*(d\Lambda^T)+(d\Lambda)\Phi^*(d\Lambda^T)\}] \ . \qquad (4.6.2.11)
\end{aligned}
$$

The second order partial matrix derivative $\nabla^2_{\Lambda_r,\Lambda_r} f(\Lambda)$ is obtained from (4.6.2.11) by applying Theorem 3.8.1 and its corollaries and finally replacing Φ^* by Φ where $\Phi = \Phi^T$. This gives

$$
\begin{aligned}
\nabla^2_{\Lambda_r,\Lambda_r} f(\Lambda) &= 2[I_{(q)}(\Phi\Lambda^T\otimes\Lambda\Phi)+I\otimes\Phi\Lambda^T\Lambda\Phi+I\otimes\Phi\Lambda^T\Lambda\Phi \\
&\quad +(\Lambda\Phi\otimes\Phi\Lambda^T)I_{(k)}+2(\Sigma-S)\otimes\Phi]
\end{aligned}
$$

$$= 4[I_{(q)}(\Phi\Lambda^T \otimes \Lambda\Phi) + I \otimes \Phi\Lambda^T\Lambda\Phi + (\Sigma - S) \otimes \Phi] \qquad (4.6.2.12)$$

since $(\Lambda\Phi \otimes \Phi\Lambda^T) I_{(k)} = I_{(q)}(\Phi\Lambda^T \otimes \Lambda\Phi)$.

The above procedure is used to obtain the remaining second order matrix derivatives which are given as follows:

$$\nabla^2_{\Phi_r, \Phi_r} f(\Phi) = 2[I_{(k)}(\Lambda^T\Lambda \otimes \Lambda^T\Lambda)] \qquad (4.6.2.13)$$

$$\nabla^2_{\Psi_d, \Psi_d} f(\Psi) = I \times\times I \qquad (4.6.2.14)$$

The procedure for obtaining mixed derivatives is as follows: We consider f as a function of only two matrix variables at a time. Then first we take the differential of f with respect to any one of the matrix variables and then the differential of the previous differential with respect to the remaining matrix variable. Theorem 3.8.1 and its corollaries then yield the required matrix derivatives. For example to obtain $\nabla^2_{\Phi_r, \Lambda_r} f$, we evaluate the differential $d_\Lambda f(\Lambda, \Phi^*)$ of f with respect to Λ and then the differential $_{\Phi^*, \Lambda} d^2 f$ of $d_\Lambda f(\Lambda, \Phi^*)$ with respect to Φ^*. Thus we have

$$_{\Phi^*, \Lambda} d^2 f(\Lambda, \Phi^*) = 2\mathrm{tr}[\{\Lambda(d\Phi^*)\Lambda^T\}\{(d\Lambda)\Phi^*\Lambda^T + \Lambda\Phi^*(d\Lambda^T)\}$$
$$+ (\Sigma^* - S)\{(d\Lambda)(d\Phi^*)\Lambda^T + \Lambda(d\Phi^*)(d\Lambda^T)\}]. \quad (4.6.2.15)$$

Therefore

$$\nabla^2_{\Phi^*_r, \Lambda_r} f(\Lambda, \Phi^*) = 2[\{\Lambda^T\Lambda(\Phi^*)^T \otimes \Lambda^T\} I_{(k)} + \Lambda^T \otimes \Lambda^T\Lambda\Phi^*$$
$$+ I_{(k)}\{\Lambda^T(\Sigma^* - S) \otimes I\} + \Lambda^T(\Sigma^* - S)^T \otimes I], \quad (4.6.2.16)$$

by applying Theorem 3.8.1 and Corollaries 3.8.1 and 3.8.3. Since $(\Phi^*)^T$ may be replaced by Φ, we obtain

$$\nabla^2_{\Phi_r,\Lambda_r} f(\Lambda,\Phi) = 2[(I_{(k)}+I)\{\Lambda^T\otimes\Lambda^T\Lambda\Phi+\Lambda^T(\Sigma-S)\otimes I\}] \qquad (4.6.2.17)$$

by using Theorem 2.5.1.

The remaining mixed derivatives may similarly be evaluated using the above procedure. These are as follows:

$$\nabla^2_{\Psi_d,\Lambda_r} f(\Lambda,\Psi) = 2N_\Psi[(I_{(k)}+I)(I\otimes\Lambda\Phi)] \qquad (4.6.2.18)$$

$$\nabla^2_{\Psi_d,\Phi_r} f(\Phi,\Psi) = 2N_\Psi[I_{(q)}(\Lambda\otimes\Lambda)]. \qquad (4.6.2.19)$$

The required Hessian matrix is then obtained by substituting the above matrices of second order and mixed derivatives in the following partitioned matrix:

$$\nabla^2_{T_r,T_r} f = \begin{bmatrix} \nabla^2_{\Lambda_r,\Lambda_r} f & \nabla^2_{\Lambda_r,\Phi_r} f & \nabla^2_{\Lambda_r,\Psi_d} f \\ \nabla^2_{\Phi_r,\Lambda_r} f & \nabla^2_{\Phi_r,\Phi_r} f & \nabla^2_{\Phi_r,\Psi_d} f \\ \nabla^2_{\Psi_d,\Lambda_r} f & \nabla^2_{\Psi_d,\Phi_r} f & \nabla^2_{\Psi_d,\Psi_d} f \end{bmatrix} \qquad (4.6.2.20)$$

where $$T_r = \begin{bmatrix} \Lambda_r \\ \Phi_r \\ \Psi_d \end{bmatrix} : (qk+k^2+q)\times 1 \qquad (4.6.2.21)$$

is a column vector of unknown parameters. Since

$$\nabla^2_{\Lambda_r, \Phi_r} f = (\nabla^2_{\Phi_r, \Lambda_r} f)^T$$

etc., the Hessian matrix $\nabla^2_{T_r, T_r} f$ defined by (4.6.2.20) is a symmetric partitioned matrix. This Hessian matrix may be used for minimizing the function $f(\Lambda, \Phi, \Psi)$ with respect to the unknown parameter matrices in the iterative procedure of Newton-Raphson and that of Fletcher & Powell (1963). A certain formulation of Fletcher-Powell method by approximate second order derivatives is given by Jöreskog (1970a, p. 240).

4.6.3 The Likelihood Model

In Section 4.6.1, we suppose that $\underline{u}$:q×1 and $\underline{\varepsilon}$:p×1 are independently distributed as multivariate normal random vectors with mean vectors zero and covariance matrices Ψ and Θ respectively. Then, since $\underline{x}$:p×1 follows a p-variate normal distribution, the density function of the observation matrix X:n×p is given by

$$f(X) = \frac{1}{(2\pi)^{\frac{pn}{2}} \, |\Sigma|^{\frac{n}{2}}} \, e^{-\frac{n}{2}\mathrm{tr}T\Sigma^{-1}} . \qquad (4.6.3.1)$$

Omitting a constant term, we write $-\frac{2}{n}$ times the logarithm of f(X) as

$$h(B, \Lambda, \Phi^*, \Psi^*, \Theta^*, \Xi) = \log_e|\Sigma^*| + \mathrm{tr}T\Sigma^{*-1} \qquad (4.6.3.2)$$

treating h to be a function of the dummy variables B, Λ, Φ^*, Ψ^*, Θ^* and Ξ.

Here we are interested in obtaining the maximum likelihood estimates $\hat{B}$, $\hat{\Lambda}$, $\hat{\Phi}$, $\hat{\Psi}$, $\hat{\Theta}$ and $\hat{\Xi}$ of the unknown parameter matrices defined in Section 4.6.1. These estimates are obtained by minimizing h with respect to the above mentioned dummy variables Jöreskog (1970a) has considered the minimization of h by an application of the modified version of the Fletcher-Powell method by using first order derivatives and an approximate value of the second order derivatives of h.

We apply very simple and general theorems from matrix differentiation to obtain the first and second order matrix derivatives of h. Then these derivatives may be used in the modified minimization procedure of Jöreskog (1970a, pp.240-241).

4.6.3.1 Maximum Likelihood Equations

For a simple factor analysis model, maximum likelihood equations were obtained by Anderson & Rubin (1956), Jöreskog (1966, 1967, 1969), Jöreskog & Lawley (1968). Bock & Bargmann (1966) obtained maximum likelihood equations when the covariance matrix in the above model may be expressed as a linear combination of known matrices. If there exists an orthogonal matrix which simultaneously diagonalizes all the above known matrices, then a set of maximum likelihood equations is given by Srivastava (1966). Further contribution to the problem of maximum likelihood estimation where the covariance matrix has a linear structure has been made by Anderson (1968). Models considered by the above authors are special cases of the general model for covariance structural analysis introduced by Jöreskog (1970a). Various specilizations and applications of the general model in the behavioural sciences are also available in the papers of Jöreskog (1970a, 1970b, 1971) where he uses the maximum likelihood method of unknown parameter estimation.

Maximum likelihood equations obtained by the above authors are based on the elementwise differentiation of a scalar function of matrices, which is a very laborious process. Matrix differentiation formulae applied by Dwyer (1967) and Gebhardt (1971) in obtaining the above equations have simplified the estimation problem to a great extent. In this sub-section we obtain the maximum likelihood equations using Theorem 3.4.3 for the general model of Jöreskog (1970a) in a much simpler way than the formulae given in Dwyer (1967) and Gebhardt (1971).

By differentiating (4.6.3.2) with respect to the dummy variables B, Λ, Φ^*, Ψ^*, Θ^* and Ξ, we obtain

$$dh = \frac{1}{|\Sigma^*|} d|\Sigma^*| + tr(dT)(\Sigma^*)^{-1} + trTd(\Sigma^*)^{-1}$$

$$= tr[\{(\Sigma^*)^{-1} - (\Sigma^*)^{-1}T(\Sigma^*)^{-1}\}d\Sigma^* + (dT)(\Sigma^*)^{-1}],$$

using $d|\Sigma^*| = |\Sigma^*| tr(\Sigma^*)^{-T} d(\Sigma^*)^T$,

$d(\Sigma^*)^{-1} = -(\Sigma^*)^{-1}(d\Sigma^*)(\Sigma^*)^{-1}$,

and some trace properties

$$= tr[W^*\{(dB)\Gamma^*B^T + B\Gamma^*(dB^T) + B[(d\Lambda)\Phi^*\Lambda^T + \Lambda(d\Phi^*)\Lambda^T + \Lambda\Phi^*(d\Lambda^T) + (d\Psi^*)]B^T + d\Theta^*\} - \frac{1}{n}\{P^T(d\Xi^T)A^T(X - A\Xi P) + (X - A\Xi P)^T A(d\Xi)P\}(\Sigma^*)^{-1}] , \tag{4.6.3.3}$$

where $W^* = (\Sigma^*)^{-1}\{I-T(\Sigma^*)^{-1}\}$, $\Gamma^* = \Lambda\Phi^*\Lambda^T+\Psi^*$. By an application of Theorem 3.4.3, the matrices of first order partial derivatives of h with respect to dummy matrices are

$$\frac{\partial h}{\partial B} = (W^*)^T B(\Gamma^*)^T + W^* B\Gamma^* \qquad (4.6.3.4)$$

$$\frac{\partial h}{\partial \Lambda} = B^T(W^*)^T B\Lambda(\Phi^*)^T + B^T W^* B\Lambda\Phi^* \qquad (4.6.3.5)$$

$$\frac{\partial h}{\partial \Phi^*} = \Lambda^T B^T (W^*)^T B\Lambda \qquad (4.6.3.6)$$

$$\frac{\partial h}{\partial \Psi^*} = B^T (W^*)^T B \qquad (4.6.3.7)$$

$$\frac{\partial h}{\partial \Theta^*} = (W^*)^T \qquad (4.6.3.8)$$

$$\frac{\partial h}{\partial \Xi} = -\frac{1}{n}A^T(X-A\Xi P)\{(\Sigma^*)^{-1}+(\Sigma^*)^{-T}\}P^T. \qquad (4.6.3.9)$$

First order matrix derivatives of h with respect to the basic parameter matrices are obtained after replacing the dummy variables by the corresponding unknown parameters and making algebraic simplification in (4.6.3.4)-(4.6.3.9). These are

$$\frac{\partial h}{\partial B} = 2WB\Gamma \qquad (4.6.3.10)$$

$$\frac{\partial h}{\partial \Lambda} = 2B^T WB\Lambda\Phi \qquad (4.6.3.11)$$

$$\frac{\partial h}{\partial \Phi} = \Lambda^T B^T WB\Lambda \qquad (4.6.3.12)$$

$$\frac{\partial h}{\partial \Psi} = B^T WB \times\times I \qquad (4.6.3.13)$$

$$\frac{\partial h}{\partial \Theta} = W \times\times I \qquad (4.6.3.14)$$

$$\frac{\partial h}{\partial \Xi} = -\frac{2}{n}A^T(X-A\Xi P)\Sigma^{-1}P^T \quad . \qquad (4.6.3.15)$$

Different expressions for the matrices (4.6.3.12), (4.6.3.13) and (4.6.3.14) were obtained by Jöreskog (1970a) without any mention of an applicable result from the Matrix Differential Calculus nor any other method of proof with a source of reference. According to our approach, for a symmetric matrix Φ, $\frac{\partial h}{\partial <\phi_{ij}>} \neq \frac{\partial h}{\partial <\phi_{ji}>}$ as matrix elements, although they are equal as scalar elements. This leads to the expression (4.6.3.12) for $\frac{\partial h}{\partial \Phi}$, which is different from that obtained by Jöreskog.

Setting (4.6.3.10)-(4.6.3.15) equal to the zero matrix, we obtain the following set of maximum likelihood equations:

$$\hat{T}\hat{\Sigma}^{-1}\hat{B}\hat{\Gamma} = \hat{B}\hat{\Gamma} \qquad (4.6.3.16)$$

$$\hat{B}^T\hat{\Sigma}^{-1}\hat{B}\hat{\Lambda}\hat{\Phi} = \hat{B}^T\hat{\Sigma}^{-1}\hat{T}\hat{\Sigma}^{-1}\hat{B}\hat{\Lambda}\hat{\Phi} \qquad (4.6.3.17)$$

$$\hat{\Lambda}^T\hat{B}^T\hat{\Sigma}^{-1}\hat{B}\hat{\Lambda} = \hat{\Lambda}^T\hat{B}^T\hat{\Sigma}^{-1}\hat{T}\hat{\Sigma}^{-1}\hat{B}\hat{\Lambda} \qquad (4.6.3.18)$$

$$\hat{B}^T\hat{\Sigma}^{-1}\hat{B}\times\times I = \hat{B}^T\hat{\Sigma}^{-1}\hat{T}\hat{\Sigma}^{-1}\hat{B}\times\times I \qquad (4.6.3.19)$$

$$\hat{\Sigma}^{-1}\times\times I = \hat{\Sigma}^{-1}\hat{T}\hat{\Sigma}^{-1}\times\times I \qquad (4.6.3.20)$$

$$A^TX\hat{\Sigma}^{-1}P^T = A^TA\hat{\Xi}P\hat{\Sigma}^{-1}P^T \quad . \qquad (4.6.3.21)$$

Maximum likelihood estimates obtained by the above equations may be improved by an iterative procedure. This uses first order matrix derivatives (4.6.3.10)-(4.6.3.15) and a Hessian matrix which we obtain in the following sub-section.

Equation (4.6.3.16) may be used to simplify the equation (4.6.3.20) in the following way: From the general model, we have

$$\hat{\Theta} = \hat{\Sigma}-\hat{B}(\hat{\Lambda}\hat{\Phi}\hat{\Lambda}^T+\hat{\Psi})\hat{B}^T$$

$$= \hat{\Sigma}-\hat{B}\hat{\Gamma}\hat{B}^T \ . \qquad (4.6.3.22)$$

Since $\hat{\Theta}$ is diagonal, pre- and post-multiplication of equation (4.6.3.20) by $\hat{\Theta}$ yields

$$(\hat{\Sigma}-\hat{B}\hat{\Gamma}\hat{B}^T)\hat{\Sigma}^{-1}(\hat{\Sigma}-\hat{B}\hat{\Gamma}\hat{B}^T)\times\times I = (\hat{\Sigma}-\hat{B}\hat{\Gamma}\hat{B}^T)\hat{\Sigma}^{-1}\hat{T}\hat{\Sigma}^{-1}(\hat{\Sigma}-\hat{B}\hat{\Gamma}\hat{B}^T)\times\times I$$

or

$$(\hat{\Sigma}+\hat{B}\hat{\Gamma}\hat{B}^T\hat{\Sigma}^{-1}\hat{B}\hat{\Gamma}\hat{B}^T)\times\times I = (\hat{T}+\hat{B}\hat{\Gamma}\hat{B}^T\hat{\Sigma}^{-1}\hat{T}\hat{\Sigma}^{-1}\hat{B}\hat{\Gamma}\hat{B}^T)\times\times I$$

or

$$\hat{\Sigma}\times\times I = \hat{T}\times\times I \qquad (4.6.3.20)'$$

using (4.6.3.16). Similarly (4.6.3.16) may be simplified as

$$\hat{T}\hat{\Theta}^{-1}\hat{B}\hat{\Gamma} = \hat{B}\hat{\Gamma}(I+\hat{B}^T\hat{\Theta}^{-1}\hat{B}\hat{\Gamma}), \qquad (4.6.3.16)'$$

which makes use of the formula

$$(B\Gamma B^T+\Theta)^{-1}B\Gamma = \Theta^{-1}B\Gamma(I+B^T\Theta^{-1}B\Gamma)^{-1}. \qquad (4.6.3.23)$$

Formulae (4.6.3.16)' and (4.6.3.20)' are very useful for computational purposes. These have some special cases, see Morrison (1967, p. 267, equations (18) and (15)). The matrix identity (4.6.3.23) makes the computation easier because this expresses the inverse of a general non-singular matrix Σ in terms of inverses of a diagonal matrix Θ and a matrix $(I+B^T\Theta^{-1}B\Gamma)$ of lower dimension. A simpler version of this identity is given by Morrison (1967, p. 267, equation (16)).

The maximum likelihood factor analysis model considered by Gebhardt (1971) may be obtained from the general model with

$\Xi = 0$ and $\Psi = I$ (in his case $B\Lambda = F$, $B = G$ and $\Theta = V$). If, in the general model, $\Xi = 0$, $B = I$ and $\Theta = 0$ then it reduces to the model discussed by Jöreskog (1966, 1967, 1969), Lawley (1967), Jöreskog & Lawley (1968) and Mulaik (1971). If, in addition, $\Phi = I$, we have the case discussed by Morrison (1967). Our approach for obtaining the likelihood equations is more general than those followed by the above authors. Besides our generality in applying various matrix formulae, we have discussed a very general model for covariance structure.

4.6.3.2 <u>The Hessian Matrix for the Likelihood Model</u>

The basic procedure for evaluating the Hessian matrix for the likelihood function is the same as that for the minres function given in Section 4.6.2.2.

We differentiate twice the function h with respect to the matrix parameter B, which gives

$$\begin{aligned} d^2h(B) &= tr[(dW^*)\{(dB)\Gamma^*B^T+B\Gamma^*(dB^T)\}+2W^*(dB)\Gamma^*(dB^T)] \\ &= tr[-(\Sigma^*)^{-1}\{(dB)\Gamma^*B^TW^*(dB)\Gamma^*B^T+(dB)\Gamma^*B^TW^*B\Gamma^*(dB^T) \\ &\quad +B\Gamma^*(dB^T)W^*(dB)\Gamma^*B^T+B\Gamma^*(dB^T)W^*B\Gamma^*(dB^T)\} \\ &\quad +U^*\{(dB)\Gamma^*B^T(\Sigma^*)^{-1}(dB)\Gamma^*B^T+(dB)\Gamma^*B^T(\Sigma^*)^{-1}B\Gamma^*(dB^T) \\ &\quad +B\Gamma^*(dB^T)(\Sigma^*)^{-1}(dB)\Gamma^*B^T+B\Gamma^*(dB^T)(\Sigma^*)^{-1}B\Gamma^*(dB^T)\} \\ &\quad +2W^*(dB)\Gamma^*(dB^T)], \end{aligned} \tag{4.6.3.24}$$

where $U^* = (\Sigma^*)^{-1}T(\Sigma^*)^{-1}$ and W^*, Γ^* are as before. Applying Theorem 3.8.1 and Corollaries 3.8.2, 3.8.4 and 3.8.6, we obtain

$$\nabla^2_{B_r,B_r} h(B) = -[I_{(p)}\{\Gamma^* B^T(\Sigma^*)^{-1} \otimes (W^*)^T B(\Gamma^*)^T\}$$
$$+(\Sigma^*)^{-1} \otimes (\Gamma^*)^T B^T (W^*)^T B(\Gamma^*)^T + (W^*)^T \otimes \Gamma^* B^T (\Sigma^*)^{-1} B\Gamma^*$$
$$+\{(\Sigma^*)^{-1} B\Gamma^* \otimes (\Gamma^*)^T B^T (W^*)^T\} I_{(q)}]$$
$$+I_{(p)}\{\Gamma^* B^T U^* \otimes (\Sigma^*)^{-T} B(\Gamma^*)^T\} + U^* \otimes (\Gamma^*)^T B^T (\Sigma^*)^{-T} B(\Gamma^*)^T$$
$$+(\Sigma^*)^{-T} \otimes \Gamma^* B^T U^* B\Gamma^* + \{U^* B\Gamma^* \otimes (\Gamma^*)^T B^T (\Sigma^*)^{-T}\} I_{(q)}$$
$$+2W^* \otimes (\Gamma^*)^T \; . \qquad (4.6.3.25)$$

Noting that $(W^*)^T = W$, $(\Gamma^*)^T = \Gamma$, $U^* = U$, and suppressing all the stars, we get after some simplification

$$\nabla^2_{B_r,B_r} h(B) = -\{I_{(p)}(\Gamma B^T \Sigma^{-1} \otimes W B\Gamma) + \Sigma^{-1} \otimes \Gamma B^T W B\Gamma + W \otimes \Gamma B^T \Sigma^{-1} B\Gamma$$
$$+(\Sigma^{-1} B\Gamma \otimes \Gamma B^T W) I_{(q)}\} + I_{(p)}(\Gamma B^T U \otimes \Sigma^{-1} B\Gamma) + U \otimes \Gamma B^T \Sigma^{-1} B\Gamma$$
$$+\Sigma^{-1} \otimes \Gamma B^T U B\Gamma + (U B\Gamma \otimes \Gamma B^T \Sigma^{-1}) I_{(q)} + 2W \otimes \Gamma$$
$$= -2[I_{(p)}(\Gamma B^T \Sigma^{-1} \otimes W B\Gamma) + \Sigma^{-1} \otimes \Gamma B^T W B\Gamma - \Sigma^{-1} S \Sigma^{-1} \otimes \Gamma B^T \Sigma^{-1} B\Gamma$$
$$-I_{(p)}(\Gamma B^T U \otimes \Sigma^{-1} B\Gamma) - W \otimes \Gamma] . \qquad (4.6.3.26)$$

Proceeding similarly, we obtain the following:

$$\nabla^2_{\Lambda_r,\Lambda_r} h(\Lambda) = -2[I_{(q)}(\Phi\Lambda^T B^T \Sigma^{-1} B \otimes B^T W B\Lambda\Phi) + B^T \Sigma^{-1} B \otimes \Phi\Lambda^T B^T W B\Lambda\Phi$$
$$-B^T U B \otimes \Phi\Lambda^T B^T \Sigma^{-1} B\Lambda\Phi - I_{(q)}(\Phi\Lambda^T B^T U B \otimes B^T \Sigma^{-1} B\Lambda\Phi) - B^T W B \otimes \Phi]$$
$$(4.6.3.27)$$

$$\nabla^2_{\Phi_r,\Phi_r} h(\Phi) = -I_{(k)}[\Lambda^T B^T \Sigma^{-1} B\Lambda \otimes \Lambda^T B^T W B\Lambda + \Lambda^T B^T U B\Lambda \otimes \Lambda^T B^T \Sigma^{-1} B\Lambda] \quad (4.6.3.28)$$

$$\nabla^2_{\Psi_d,\Psi_d} h(\Psi) = -[B^T \Sigma^{-1} B \times\times B^T (W-U) B] \quad (4.6.3.29)$$

$$\nabla^2_{\Theta_d,\Theta_d} h(\Theta) = -[\Sigma^{-1} \times\times (W-U)] \quad (4.6.3.30)$$

$$\nabla^2_{\Xi_r,\Xi_r} h(\Xi) = \frac{2}{n}(A^T A \otimes P\Sigma^{-1} P^T) \quad (4.6.3.31)$$

$$\nabla^2_{B_r,\Lambda_r} h(B,\Lambda) = 2[-(WB\Lambda\Phi \otimes \Gamma B^T \Sigma^{-1} B) I_{(k)} + (\Sigma^{-1} B\Lambda\Phi \otimes \Gamma B^T U B) I_{(k)}$$

$$-WB \otimes \Gamma B^T \Sigma^{-1} B\Lambda\Phi + \Sigma^{-1} B \otimes \Gamma B^T U B\Lambda\Phi$$

$$+(WB\Lambda\Phi \otimes I) I_{(k)} + WB \otimes \Lambda\Phi] \quad (4.6.3.32)$$

$$\nabla^2_{B_r,\Phi_r} h(B,\Phi) = [\{UB\Lambda \otimes \Gamma B^T \Sigma^{-1} B\Lambda + WB\Lambda \otimes \Lambda - \Sigma^{-1} B\Lambda \otimes \Gamma B^T W B\Lambda\} I_{(k)}$$

$$-WB\Lambda \otimes (\Gamma B^T \Sigma^{-1} B - I)\Lambda + \Sigma^{-1} B\Lambda \otimes \Gamma B^T U B\Lambda] . \quad (4.6.3.33)$$

The remaining mixed partial matrix derivatives may similarly be evaluated.

Let

$$\Omega_r = \begin{bmatrix} B_r \\ \Lambda_r \\ \Phi_r \\ \Psi_d \\ \Theta_d \\ \Xi_r \end{bmatrix} : (pq+qk+k^2+q+p+gh) \times 1. \quad (4.6.3.34)$$

The Hessian matrix $\nabla^2_{\Omega_r,\Omega_r} h(\Omega)$ is then obtained by assembling the above matrices in the form of the following symmetric partitioned matrix:

$$\nabla^2_{\Omega_r,\Omega_r} h = \begin{bmatrix} \nabla^2_{B_r,B_r} h & \nabla^2_{B_r,\Lambda_r} h & \nabla^2_{B_r,\Phi_r} h & \nabla^2_{B_r,\Psi_d} h & \nabla^2_{B_r,\Theta_d} h & \nabla^2_{B_r,\Xi_r} h \\ \nabla^2_{\Lambda_r,B_r} h & \nabla^2_{\Lambda_r,\Lambda_r} h & \nabla^2_{\Lambda_r,\Phi_r} h & \nabla^2_{\Lambda_r,\Psi_d} h & \nabla^2_{\Lambda_r,\Theta_d} h & \nabla^2_{\Lambda_r,\Xi_r} h \\ \nabla^2_{\Phi_r,B_r} h & \nabla^2_{\Phi_r,\Lambda_r} h & \nabla^2_{\Phi_r,\Phi_r} h & \nabla^2_{\Phi_r,\Psi_d} h & \nabla^2_{\Phi_r,\Theta_d} h & \nabla^2_{\Phi_r,\Xi_r} h \\ \nabla^2_{\Psi_d,B_r} h & \nabla^2_{\Psi_d,\Lambda_r} h & \nabla^2_{\Psi_d,\Phi_r} h & \nabla^2_{\Psi_d,\Psi_d} h & \nabla^2_{\Psi_d,\Theta_d} h & \nabla^2_{\Psi_d,\Xi_r} h \\ \nabla^2_{\Theta_d,B_r} h & \nabla^2_{\Theta_d,\Lambda_r} h & \nabla^2_{\Theta_d,\Phi_r} h & \nabla^2_{\Theta_d,\Psi_d} h & \nabla^2_{\Theta_d,\Theta_d} h & \nabla^2_{\Theta_d,\Xi_r} h \\ \nabla^2_{\Xi_r,B_r} h & \nabla^2_{\Xi_r,\Lambda_r} h & \nabla^2_{\Xi_r,\Phi_r} h & \nabla^2_{\Xi_r,\Psi_d} h & \nabla^2_{\Xi_r,\Theta_d} h & \nabla^2_{\Xi_r,\Xi_r} h \end{bmatrix} . \tag{4.6.3.35}$$

4.6.4 A Generalization of Howe's Function

In this section we show that under certain distribution-free assumptions a set of estimating equations for B, Λ, Φ, Ψ and Θ of the structural form (4.6.1.3) may be obtained. For the models (4.6.1.1) and (4.6.1.2), the only supposition we make is that $Var(\underline{x}) = \Sigma$, $Var(\underline{z}) = \Phi$ and $Var(\underline{u}) = \Psi$:diagonal matrix, are well-defined covariance matrices. Then the following hypothesis gives rise to the more general structural form of Σ given by (4.6.1.3): Does there exist a random vector $\underline{z}$ with non-diagonal covariance matrix Φ such that the partial correlations of the elements of the observable random vector $\underline{x}$ with all elements of $\underline{z}$ held constant is 0? With

$$\Theta = [E(\underline{x}-B\underline{y})(\underline{x}-B\underline{y})^T]\times\times I \ , \qquad (4.6.4.1)$$

a matrix formulation of this hypothesis is

$$\Theta^{-\frac{1}{2}}[E(\underline{x}-B\underline{y})(\underline{x}-B\underline{y})^T]\Theta^{-\frac{1}{2}} = I \qquad (4.6.4.2)$$

since $\rho_{x_i,x_j;\underline{z}} = 0$ for all i, j.

Since

$$E(\underline{x}-B\underline{y})(\underline{x}-B\underline{y})^T = \Sigma-B(\Lambda\Phi\Lambda^T+\Psi)B^T \ ,$$

we have from (4.6.4.2)

$$\Sigma = B(\Lambda\Phi\Lambda^T+\Psi)B^T+\Theta \qquad (4.6.4.3)$$

which is the same as (4.6.1.3).

Under the above hypothesis, we may rewrite (4.6.4.2) as

$$\Theta^{-\frac{1}{2}}[\Sigma-B(\Lambda\Phi\Lambda^T+\Psi)B^T]\Theta^{-\frac{1}{2}} = I \ . \qquad (4.6.4.4)$$

If in (4.6.4.4), we replace Σ by the sample covariance matrix S, then minimizing all the sample partial correlation coefficients simultaneously is equivalent to maximizing the following determinant

$$|\Theta^{-\frac{1}{2}}[S-B(\Lambda\Phi\Lambda^T+\Psi)B^T]\Theta^{-\frac{1}{2}}| \ . \qquad (4.6.4.5)$$

Equation (4.6.4.5) may be treated as a scalar function of the dummy variables B, Λ, Φ^*, Ψ^* and Θ^*. It is represented more simply as

$$g(B, \Lambda, \Phi^*, \Psi^*, \Theta^*) = \frac{|S-B(\Lambda\Phi^*\Lambda^T+\Psi^*)B^T|}{|\Theta^*|} \ . \qquad (4.6.4.6)$$

We call function (4.6.4.6) the generalized Howe's function. A very particular case of this function has been discussed by Morrison (1967, pp. 287-288). Our B, Λ, Θ and $\underline{u}$ are his Λ, I, Ψ and $\underline{O}$ respectively.

4.6.4.1 Minimum Partial Correlation Coefficients Estimates

These estimates are obtained by maximizing the generalized Howe's function (4.6.4.6) with respect to the unknown parameter matrices. Differentiation of g with respect to the dummy variables B, Λ, Φ^*, Ψ^* and Θ^*, yields

$$
\begin{aligned}
dg(B,\ \Lambda,\ \Phi^*,\ \Psi^*) &= \frac{1}{|\Theta^*|}\, d|S-B(\Lambda\Phi^*\Lambda^T+\Psi^*)B^T| \\
&\quad +|S-B(\Lambda\Phi^*\Lambda^T+\Psi^*)B^T|\, d(\frac{1}{|\Theta^*|}) \\
&= \frac{|S-B\Gamma^*B^T|}{|\Theta^*|}\, tr[-\{S-B\Gamma^*B^T\}^{-T} d\{B(\Gamma^*)^T B^T\}] \\
&\quad -\frac{|S-B\Gamma^*B^T|}{|\Theta^*|}\, tr(\Theta^*)^{-T} d(\Theta^*)^T \\
&= \frac{|S-B\Gamma^*B^T|}{|\Theta^*|}\, tr\Big[\{(\Theta^*)^{-T}-(S-B\Gamma^*B^T)^{-T}\} \\
&\quad [(dB)(\Gamma^*)^T B^T+B(\Gamma^*)^T(dB^T)+B\{(d\Lambda)(\Phi^*)^T\Lambda^T \\
&\quad +\Lambda d(\Phi^*)^T\Lambda^T+\Lambda(\Phi^*)^T(d\Lambda^T)+d(\Psi^*)^T\}B^T]\Big],
\end{aligned}
\tag{4.6.4.7}
$$

and

$$
dg(\Theta^*) = \frac{|S-B\Gamma^*B^T|}{|\Theta^*|}\, tr[\{(S-B\Gamma^*B^T)^{-T}-(\Theta^*)^{-T}\}d(\Theta^*)^T]. \tag{4.6.4.8}
$$

Applying Theorem 3.4.3 and making algebraic simplifications similar to those in Section 4.6.2.1, we obtain

$$\frac{\partial g}{\partial B} = \frac{|S-B\Gamma B^T|}{|\Theta|}\{\Theta^{-1}-(S-B\Gamma B^T)^{-1}\}B\Gamma \tag{4.6.4.9}$$

$$\frac{\partial g}{\partial \Lambda} = \frac{|S-B\Gamma B^T|}{|\Theta|} B^T\{\Theta^{-1}-(S-B\Gamma B^T)^{-1}\}B\Lambda\Phi \tag{4.6.4.10}$$

$$\frac{\partial g}{\partial \Phi} = \frac{|S-B\Gamma B^T|}{|\Theta|} \Lambda^T B^T\{\Theta^{-1}-(S-B\Gamma B^T)^{-1}\}B\Lambda \tag{4.6.4.11}$$

$$\frac{\partial g}{\partial \Psi} = \frac{|S-B\Gamma B^T|}{|\Theta|} B^T\{\Theta^{-1}-(S-B\Gamma B^T)^{-1}\}B\times\times I \tag{4.6.4.12}$$

$$\frac{\partial g}{\partial \Theta} = \frac{|S-B\Gamma B^T|}{|\Theta|} \{(S-B\Gamma B^T)^{-1}-\Theta^{-1}\} \quad . \tag{4.6.4.13}$$

The required equations for estimating the parameters B, Λ, Φ, Ψ and Θ are then obtained by setting equations (4.6.4.9)-(4.6.4.13) equal to zero. These, after some simplification, give

$$S\hat{\Theta}^{-1}\hat{B}\hat{\Gamma} = \hat{B}\hat{\Gamma}(I+\hat{B}^T\hat{\Theta}^{-1}\hat{B}\hat{\Gamma}) \tag{4.6.4.14}$$

$$\hat{B}^T(S-\hat{B}\hat{\Gamma}\hat{B}^T)^{-1}\hat{B}\hat{\Lambda}\hat{\Phi} = \hat{B}^T\hat{\Theta}^{-1}\hat{B}\hat{\Lambda}\hat{\Phi} \tag{4.6.4.15}$$

$$\hat{\Lambda}^T\hat{B}^T(S-\hat{B}\hat{\Gamma}\hat{B}^T)^{-1}\hat{B}\hat{\Lambda} = \hat{\Lambda}^T\hat{B}^T\hat{\Theta}^{-1}\hat{B}\hat{\Lambda} \tag{4.6.4.16}$$

$$\hat{B}^T(S-\hat{B}\hat{\Gamma}\hat{B}^T)^{-1}\hat{B}\times\times I = \hat{B}^T\hat{\Theta}^{-1}\hat{B}\times\times I \tag{4.6.4.17}$$

$$\hat{\Sigma}\times\times I = S\times\times I \quad . \tag{4.6.4.18}$$

The estimating equations (4.6.3.16)', (4.6.3.20)', obtained by the method of maximum likelihood, are analogous to the equations (4.6.4.14) and (4.6.4.18) respectively. The equivalence of these equations provides some justification for adopting the generalized Howe's function represented by (4.6.4.6) for estimating the unknown parameter matrices. For a simple factor analysis model, Morrison(1967, p. 289) has obtained an estimation equation using Howe's procedure. His equation is a particular case of our equation (4.6.4.14), where our (B, Γ, Θ) are his (Λ, I, Ψ). The maximization procedure followed by Morrison(1967, p. 288) is much more complicated even for a very simple case. Our procedure is very straightforward and can be used even for more general matrix functions than those considered so far. The treatment of Sections 4.6.3 and 4.6.4 is based on Tracy & Singh (1971a). In particular, Singh & Tracy (1970) discussed likelihood function and Howe's function for a factor analysis model considered by Jöreskog (1969).

4.6.5 Estimation of Unobserved Variables

Suppose that in the model (4.6.1.1) and (4.6.1.2), $\underline{y}$ and hence $\underline{z}$, $\underline{u}$ are unobserved non-random vectors. Let

$$X = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} : n \times p \qquad (4.6.5.1)$$

be an observed random matrix.

Since $\underline{z}$ and $\underline{u}$ are non-random vectors, we have for each j

$$\underline{x}_j = B\Lambda\underline{z}_j + B\underline{u}_j + \underline{e}_j \quad , \qquad (4.6.5.2)$$

where $\underline{z}_j$ and $\underline{u}_j$ may be treated as unknown parameter vectors.

If we define

$$Z = \begin{bmatrix} \underline{z}_1^T \\ \vdots \\ \underline{z}_n^T \end{bmatrix}, \quad U = \begin{bmatrix} \underline{u}_1^T \\ \vdots \\ \underline{u}_n^T \end{bmatrix} \text{ and } E = \begin{bmatrix} \underline{e}_1^T \\ \vdots \\ \underline{e}_n^T \end{bmatrix}, \tag{4.6.5.3}$$

then X may be expressed as a function of the partitioned matrix $[B\Lambda \vdots B]$ in the following way:

$$X = Z\Lambda^T B^T + UB^T + E$$

$$= [Z \vdots U] \begin{bmatrix} \Lambda^T \\ \cdot\cdot \\ I \end{bmatrix} B^T + E \ . \tag{4.6.5.4}$$

Our aim in this section is to obtain the weighted least squares estimate of $[Z \vdots U]$. This requires minimizing the function

$$f(W) = \text{tr}(X-WD)\Theta^{-1}(X-WD)^T \tag{4.6.5.5}$$

where

$$W = [Z \vdots U] \text{ and } D^T = B[\Lambda \vdots I]. \tag{4.6.5.6}$$

On differentiating (4.6.5.5) with respect to W, we get

$$df(W) = \text{tr}[-X\Theta^{-1}D^T(dW^T)-(dW)D\Theta^{-1}X^T+(dW)D\Theta^{-1}D^TW^T$$

$$+WD\Theta^{-1}D^T(dW^T)]. \tag{4.6.5.7}$$

Applying Theorem 3.4.3, we obtain

$$\frac{\partial f(W)}{\partial W} = 2(WD\Theta^{-1}D^T - X\Theta^{-1}D^T). \tag{4.6.5.8}$$

Let $\hat{\hat{W}}$ be the weighted least squares estimate of W, where

$\hat{\hat{W}}$ is obtained from known estimates of the parameters. Then setting $\frac{\partial f}{\partial W} = 0$ and assuming that $D\Theta^{-1}D^T$ is non-singular, we obtain the following estimating equation:

$$\hat{\hat{W}} = X\hat{\Theta}^{-1}\hat{D}^T(\hat{D}\hat{\Theta}^{-1}\hat{D}^T)^{-1} \quad , \tag{4.6.5.9}$$

where $\hat{\Theta}$ and $\hat{D}^T = \hat{B}[\hat{\Lambda} \vdots I]$ are known least squares estimates of Θ and D^T respectively. Since

$$D\Theta^{-1}D^T = \begin{bmatrix} \Lambda^T B^T \Theta^{-1} B\Lambda & \vdots & \Lambda^T B^T \Theta^{-1} B \\ \cdots & \cdots & \cdots \\ B^T \Theta^{-1} B\Lambda & \vdots & B^T \Theta^{-1} B \end{bmatrix} , \tag{4.6.5.10}$$

we may express (4.6.5.9) more explicitly as

$$[\hat{\hat{Z}} \vdots \hat{\hat{U}}] = X\hat{\Theta}^{-1}\hat{B}[\hat{\Lambda} \vdots I] \begin{bmatrix} \hat{\Lambda}^T \hat{B}^T \hat{\Theta}^{-1} \hat{B}\hat{\Lambda} & \vdots & \hat{\Lambda}^T \hat{B}^T \hat{\Theta}^{-1} \hat{B} \\ \cdots & \cdots & \cdots \\ \hat{B}^T \hat{\Theta}^{-1} \hat{B}\hat{\Lambda} & \vdots & \hat{B}^T \hat{\Theta}^{-1} \hat{B} \end{bmatrix}^{-1} . \tag{4.6.5.11}$$

For evaluating the inverse of the partitioned matrix (4.6.5.10) we refer to Graybill (1969, pp. 164-166).

Particular cases of this estimation problem have been treated by Anderson & Rubin (1956), Lawley & Maxwell (1963) and Morrison (1967), without using any matrix differentiation formulae.

We conclude our discussion on parameter estimation in the covariance structural analysis by claiming that the matrix derivative approach is the most efficient procedure to deal with such complicated scalar functions of matrices.

4.7 Evaluation of Jacobians Under Symmetric Matrix Transformations

Matrix derivatives with respect to matrix elements are of general application in multivariate analysis. Some of these applications occur in estimation problems, in obtaining large sample covariance matrices and in evaluating jacobians of certain matrix transformations. Under general matrix transformations $Y = F(X)$, the jacobian is simply the absolute value of the determinant of $\frac{\partial Y_r}{\partial X_r}$. For example, (i) if $Y = AX^TB$, $A{:}p\times p$, $B{:}q\times q$, then

$$\text{Jacobian} = \text{mod}\left|\frac{\partial Y_r}{\partial X_r}\right|$$

$$= \text{mod}\,|I_{(q)}(A^T \otimes B)|$$

$$= |A|^q|B|^p \text{, except for a sign, using (2.4.6) and (A.1.27)} \qquad (4.7.1)$$

and (ii) if $Y = X^{-T}$, $X{:}p\times p$, then

$$\text{Jacobian} = \text{mod}\,|-I_{(p)}(X^{-1} \otimes X^{-T})|$$

$$= |X|^{-2p} \text{, except for a sign.} \qquad (4.7.2)$$

Matrix derivatives with respect to scalar elements are also useful in the above problems, but this approach is very complicated because these are not general matrix derivatives. However, for evaluating jacobians, these are the most desirable matrix derivatives, especially where matrix transformations involve known equality relationships. For example, if

(i) $Y = AX^TB$, (ii) $Y = X^{-T}$, with $X = X^T$ in both cases, then (4.7.1) and (4.7.2.) are not the appropriate jacobians.

The aim of this section is to evaluate the jacobians of some special matrix transformations involving symmetric matrices by applying matrix derivative results from Section 3.11. Some of these jacobians have been evaluated by Dwyer (1967), using matrix derivatives, and by Olkin & Sampson (1969), using a functional equation induced on a vector space of symmetric matrices. The generality of our approach lies in the fact that only one modified matrix derivative formula, given by Corollary 3.11.2.1, is enough to evaluate various jacobians for a class of symmetric matrix transformations. However, from (3.11.3.9), it is obvious that either Corollary 3.11.2.1 or Theorem 3.11.3.3 may be used for evaluating such jacobians.

We denote the jacobian of a special matrix transformation $Y = F(X)$, $X = X^T$, by

$$J(X \to Y) = \mathrm{mod}\left|\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}}\right| . \tag{4.7.3}$$

We use this particular form of matrix derivative because it is easier to handle than the one discussed in Section 3.11.3.

We require the following matrix result given by Hsu (1953, p. 41):

Theorem 4.7.1 Every non-singular matrix may be expressed as the product of a finite number of matrices of either of the following two types:

(i) a diagonal matrix whose diagonal elements are 1 with the exception of one, which is, say, a;

(ii) a matrix whose diagonal elements are 1 and all but one of whose non-diagonal elements are zero.

Example 4.7.1 If

$$Y = AXA^T , \tag{4.7.4}$$

where $A:p\times p$ is a non-singular matrix, and $X = X^T$, then

$$J(X \to Y) = |A|^{p+1}. \tag{4.7.5}$$

To get (4.7.5), taking the differential of (4.7.4), we obtain

$$dY = A(dX)A^T \tag{4.7.6}$$

$$\Longrightarrow dY_r = (A\otimes A)dX_r \text{ , using (2.6.3.13).} \tag{4.7.7}$$

Now applying Corollary 3.11.2.1 to (4.7.7), we get

$$\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}} = [M_Y(A\otimes A)N_X]^T . \tag{4.7.8}$$

Hence

$$J(X \to Y) = |M_Y(A\otimes A)N_X| . \tag{4.7.9}$$

To evaluate (4.7.9) we apply Theorem 4.7.1. Without any loss of generality we may assume $A = \text{diag}(a, 1,\ldots, 1)$. Then

$$A\otimes A = \text{diag}(\underset{\overset{\uparrow}{1}}{a^2}, \underset{\overset{\uparrow}{2}}{a},\ldots, \underset{\overset{\uparrow}{p}}{a}, \underset{\overset{\uparrow}{p+1}}{a}, 1,\ldots, 1, \underset{\overset{\uparrow}{2p+1}}{a}, 1,\ldots, 1,\ldots, \underset{\underset{(p-1)p+1}{\downarrow}}{a}, 1,\ldots, 1)$$

and $M_Y(A\otimes A)N_X$ is a lower triangular matrix whose diagonal is given by

$$\begin{matrix} (a^2, & a, \ldots, & a, & 1, \ldots, & 1). \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 2 & p & p+1 & p(p+1) \end{matrix}$$

Hence

$$|M_Y(A \otimes A)N_X| = |A|^{p+1}$$

which yields (4.7.5).

In particular, if in (4.7.4), A is a non-singular upper triangular matrix ($a_{ij} = 0$, $i > j$), then this transformation is needed in connection with the Wishart distribution [see Anderson (1958, p. 157)]. Evaluation of (4.7.9) for this particular case is much simpler and has been considered by Tracy & Singh (1970a).

Using the above example, we obtain the following jacobian of an inverse matrix transformation:

Example 4.7.2 If

$$Y = X^{-1} , \tag{4.7.10}$$

where $X = X^T$:p×p, then

$$J(X \to Y) = |X|^{-(p+1)} . \tag{4.7.11}$$

Since YX = I, we have

$$(dY)X+Y(dX) = 0$$

$$\Longrightarrow \quad dY = -X^{-1}(dX)X^{-1}$$

$$\Longrightarrow \quad dY_r = -(X^{-1} \otimes X^{-T})dX_r . \tag{4.7.12}$$

Expression (4.7.12) is of the form (4.7.7) and hence to obtain the jacobian of (4.7.10), we replace A by X^{-1} in (4.7.5) and get (4.7.11).

Dwyer (1967) and Olkin & Sampson (1969) evaluated (4.7.5) and (4.7.11) using different approaches. Our approach brings out certain advantages over that of Dwyer (1967) by making use of Kronecker product and an interesting property of a non-singular matrix given by Hsu (1953).

Example 4.7.3 The jacobian of the matrix transformation

$$Y = XAX \ , \tag{4.7.13}$$

where $X = X^T:p\times p$, and $A = A^T$ is

$$J(X \rightarrow Y) = \prod_{i\leqslant j}(\lambda_i+\lambda_j) \tag{4.7.14}$$

where λ_i are the characteristic roots of AX.

Taking the differential of (4.7.13) leads, after some simplification, to

$$\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}} = N_X^T[I\otimes AX+AX\otimes I]M_Y^T \ . \tag{4.7.15}$$

In obtaining the jacobian of (4.7.13), we take $AX = \text{diag}(\lambda_1,\ldots, \lambda_p)$, and consequently the required determinant of (4.7.15) is $\prod_{i\leqslant j}(\lambda_i+\lambda_j)$.

If, in particular, $Y = X^2$, $X = X^T:p\times p$, then the required jacobian may be obtained from the above example and is given by

$$J(X \rightarrow X^2) = \prod_{i\leqslant j}(\lambda_i+\lambda_j) \ , \tag{4.7.16}$$

where λ_i are the characteristic roots of X.

Olkin & Sampson (1969) obtained (4.7.14) by using a chain of matrix transformations and multiplying their jacobians. Also they provided a very complicated proof for (4.7.16) by solving the functional equations for a chain of suitable matrix transformations.

Example 4.7.4 The jacobian of the matrix transformation

$$Y = AXA^T+B^TXB \ , \tag{4.7.17}$$

where $X = X^T$:p×p matrix, is given by

$$J(X \to Y) = \prod_{i\leqslant j} (\lambda_i\lambda_j+\mu_i\mu_j) \ , \tag{4.7.18}$$

where λ_i and μ_j are the characteristic roots of A and B respectively.

Taking differential of (4.7.17) and using (2.6.3.13) and Corollary 3.11.2.1, we obtain

$$\frac{\partial Y_r^{\#}}{\partial X_{(r)}^{\#}} = N_X^T[A^T\otimes A^T+B\otimes B]M_Y^T \ , \tag{4.6.19}$$

where M_Y and N_X are $\frac{p(p+1)}{2}\times p^2$ and $p^2\times\frac{p(p+1)}{2}$ appropriate matrices respectively.

From a particular case of (A.1.28), $\lambda_i\lambda_j$ and $\mu_i\mu_j$; $i,j = 1, 2,\ldots, p$; are the characteristic roots of $A^T\otimes A^T$ and $B\otimes B$ respectively.

Hence the result follows from Example 4.7.3.

4.8 Miscellaneous

In this section we consider a number of important miscellaneous applications of Chapters II and III in multivariate analysis.

4.8.1 Some Properties of the Matrix Product Defined by Khatri & Rao (1968) and Extended by Khatri (1971)

The proof of the following result is straightforward:

Theorem 4.8.1.1 Let

$$A = (A_{.1},\dots, A_{.q}):m\times q;\ B = (B_{.1},\dots, B_{.q}):n\times q,$$

then

$$I_{(n)}(A\odot B) = B\odot A \qquad (4.8.1.1)$$

$$I_{(m)}(B\odot A) = A\odot B \qquad (4.8.1.2)$$

where $A\odot B$ is as in Definition 2.7.2 defined by Khatri & Rao (1968), and $I_{(k)}$ is as in Definition 2.4.1.

Definition 4.8.1.1 Let $I: \sum_{i=1}^{r} q_i n_i \times \sum_{i=1}^{r} q_i n_i$ be partitioned as

$$I = \text{Diag}[\overset{q_1n_1}{I} : \overset{q_2n_2}{I} : \overset{\cdots}{\dots} : \overset{q_rn_r}{I}]. \qquad (4.8.1.3)$$

Then we define

$$I_{\langle q_i,i=1,2,\dots,r;n\rangle} = \text{Diag}[\overset{q_1n_1}{I_{(q_1)}} : \overset{q_2n_2}{I_{(q_2)}} : \overset{\cdots}{\dots} : \overset{q_rn_r}{I_{(q_r)}}] \qquad (4.8.1.4)$$

and

$$I_{\langle n_i,i=1,2,\dots,r;q\rangle} = \text{Diag}[\overset{q_1n_1}{I_{(n_1)}} : \overset{q_2n_2}{I_{(n_2)}} : \overset{\cdots}{\dots} : \overset{q_rn_r}{I_{(n_r)}}], \qquad (4.8.1.5)$$

where $\sum_{i=1}^{r} n_i = n$, $\sum_{i=1}^{r} q_i = q$.

Some interesting properties of the auxiliary operators defined by (4.8.1.4) and (4.8.1.5) are given by the following

Theorem 4.8.1.2

(i) $I_{<q_i,i=1,2,...,r;n>} = (I_{<n_i,i=1,2,...,r;q>})^T$ (4.8.1.6)

(ii) $(I_{<q_i,i=1,2,...,r;n>})(I_{<n_i,i=1,2,...,r;q>}) = I$ (4.8.1.7)

(iii) $(I_{<n_i,i=1,2,...,r;q>})(I_{<q_i,i=1,2,...,r;n>}) = I$ (4.8.1.8)

(iv) $I_{<q_i,i=1,2,...,r;n>} = I_{<n_i,i=1,2,...,r;q>}$

if and only if $q = n$ and $q_i = n_i$ for all i (4.8.1.9)

(v) $I_{<n_i,i=1,2,...,r;q>}+I_{<q_i,i=1,2,...,r;n>}$ is a symmetric partitioned matrix (4.8.1.10)

(vi) $|I_{<n_i,i=1,2,...,r;q>}| = |I_{<q_i,i=1,2,...,r;n>}| = \pm 1$. (4.8.1.11)

The verification of all the above properties is straightforward.

We need Definition 4.8.1.1 and Theorem 4.8.1.2 to establish certain connections between the auxiliary operators $I_{(k)}$ introduced by Tracy & Dwyer (1969) and the matrix product ⓔ defined by Khatri (1971). These are given by the following

Theorem 4.8.1.3 Let A:m×n, B:p×q matrices which are partitioned as follows:

$$A = \overset{\quad n_1 \quad n_2 \quad \cdots \quad n_r}{[A^1 \vdots\ A^2 \vdots\ \cdots \vdots\ A^r]}, \quad B = \overset{\quad q_1 \quad q_2 \quad \cdots \quad q_r}{[B^1 \vdots\ B^2 \vdots\ \cdots \vdots\ B^r]};$$

then

(i) $I_{(p)}(A ⓔ B) = (B ⓔ A)I_{<q_i,i=1,2,...,r;n>}$ (4.8.1.12)

(ii) $I_{(m)}(B \circledcirc A) = (A \circledcirc B)I_{<n_i, i=1,2,\ldots,r;q>}$ (4.8.1.13)

(iii) $I_{(p)}(A \circledcirc B)I_{<n_i, i=1,2,\ldots,r;q>} = B \circledcirc A$, (4.8.1.14)

where $A \circledcirc B$ is defined by Khatri (1971); see Definition 2.7.3.

Proof: (i) The i-th column block of $I_{(p)}(A \circledcirc B)$ is given by

$$I_{(p)}(A^i \otimes B^i) = (B^i \otimes A^i)I_{(q_i)} \;. \qquad (4.8.1.15)$$

The right-hand side of expression (4.8.1.15) is the i-th column block of $(B \circledcirc A)I_{<q_i, i=1,2,\ldots,r;n>}$. Hence the result.

The proof of (ii) is analogous to (i).

(iii) Applying (4.8.1.7) and (4.8.1.12), we obtain

$$I_{(p)}(A \circledcirc B)I_{<n_i, i=1,2,\ldots,r;q>}$$

$$= (B \circledcirc A)(I_{<q_i, i=1,2,\ldots,r;n>})(I_{<n_i, i=1,2,\ldots,r;q>})$$

$$= B \circledcirc A.$$

Hence the theorem is proved.

Obviously

$$I_{(m)}(B \circledcirc A)I_{<q_i, i=1,2,\ldots,r;n>} = A \circledcirc B \;. \qquad (4.8.1.16)$$

Here we may point out that just as the matrix product $\odot$ is a special case of $\circledcirc$, it is immediate that Theorem 4.8.1.1 is a special case of Theorem 4.8.1.3. Further, since $\circledcirc$ is a special case of our partitioned matrix product $\textcircled{\pi}$, it is observed that Theorem 4.8.1.3 is a special case of Theorem 2.9.1. If we have a multivariate model $Y = AXB^T$, where A and B are as in Theorems 4.8.1.1 or 4.8.1.3, and X is a diagonal or a block diagonal matrix, then the results of this subsection are very useful.

4.8.2 Estimation of Scalar Coefficients for Covariance Matrices with Linear Structure

Complications may arise even if we are required to differentiate a scalar function of matrices with respect to scalar variables. For example, in the analysis of covariance structures, both the positive definite matrix and its inverse may be assumed to be certain linear combinations of known, linearly independent symmetric matrices involving unknown scalar coefficients. The maximum likelihood estimation of these unknown coefficients requires the differentiation of the log-likelihood function with respect to these dummy variables.

In this subsection we obtain a set of likelihood equations and a Hessian matrix concerning likelihood functions involving a positive definite covariance matrix and its inverse, which are represented as linear combinations of known matrices.

Such an estimation problem for a positive definite covariance matrix with a linear structure is discussed by Srivastava (1966) and Bock & Bargmann (1966). Anderson (1968) gives a simpler treatment for the problem considered by the above authors and also for estimating the unknown scalar coefficients when the inverse of the covariance matrix is a linear combination of known matrices. However, the procedure for obtaining the likelihood equation followed by the above authors may be considerably simplified by using particular cases of our matrix differentiation formulae. Here we discuss these simplifications in the light of the work of Anderson (1968).

Omitting a constant term, we consider the minimization with respect to σ_i, i = 0, 1, ..., m, of

$$g(\sigma_0, \sigma_1, \ldots, \sigma_m) = \log|\Sigma| + \mathrm{tr}S\Sigma^{-1} \qquad (4.8.2.1)$$

where S is a sample covariance matrix; $\Sigma = \sum_{i=0}^{m} \sigma_i G_i$, G_i are known and linearly independent symmetric matrices, and the σ_i are unknown parameters.

Differentiating (4.8.2.1) with respect to these parameters, we obtain

$$dg = \mathrm{tr}(\Sigma^T)^{-1}\left(\sum_{i=0}^{m} G_i d\sigma_i\right) - \mathrm{tr}S\Sigma^{-1}\left(\sum_{i=0}^{m} G_i d\sigma_i\right)\Sigma^{-1} \qquad (4.8.2.2)$$

$$= \sum_{i=0}^{m} \mathrm{tr}[\Sigma^{-1}G_i d\sigma_i - S\Sigma^{-1}G_i(d\sigma_i)\Sigma^{-1}] \ . \qquad (4.8.2.3)$$

From (4.8.2.3), using a special case of Theorem 3.4.3, we get

$$\frac{\partial g}{\partial \sigma_i} = \mathrm{tr}\Sigma^{-1}G_i(I - \Sigma^{-1}S), \ i = 0, 1, \ldots, m. \qquad (4.8.2.4)$$

Hence the required likelihood equations are

$$\mathrm{tr}\hat{\Sigma}^{-1}G_i = \mathrm{tr}\hat{\Sigma}^{-1}G_i\hat{\Sigma}^{-1}S \ , \ i = 0, 1, \ldots, m, \qquad (4.8.2.5)$$

where $\hat{\Sigma} = \sum_{i=0}^{m} \hat{\sigma}_i G_i$ and $\hat{\sigma}_i$ are solutions of (4.8.2.5).

Anderson (1968, pp. 57-58) obtained (4.8.2.5) using scalar differentiation which is quite involved. In his case the final result is obtained by using the following formulae:

$$\frac{\partial \sigma_{k\ell}}{\partial \sigma_i} = g_{k\ell}^{(i)} \ , \quad k,\ell = 1, 2, \ldots, p \ ,$$

$$\frac{\partial \Sigma^{-1}}{\partial \sigma_i} = -\Sigma^{-1} G_i \Sigma^{-1}$$

and

$$\frac{\partial \log|\Sigma|}{\partial \sigma_i} = \sum_{k=1}^{p} \sum_{\ell=1}^{p} \frac{\partial \log|\Sigma|}{\partial \sigma_{k\ell}} \frac{\partial \sigma_{k\ell}}{\partial \sigma_i} = \sum_{k=1}^{p} \sum_{\ell=1}^{p} \left(\frac{\text{cof } \sigma_{k\ell}}{|\Sigma|}\right) g_{k\ell}^{(i)}$$

$$= \sum_{k=1}^{p} \sum_{\ell=1}^{p} \sigma^{k\ell} g_{k\ell}^{(i)} \ .$$

Further differentiation of (4.8.2.1) first with respect to σ_i and then σ_j gives

$$\begin{aligned} d^2_{\sigma_j,\sigma_i} g &= -\mathrm{tr}(\Sigma^T)^{-1}(G_j d\sigma_j)(\Sigma^T)^{-1}(G_i d\sigma_i) \\ &+\mathrm{tr}S\Sigma^{-1}(G_j d\sigma_j)\Sigma^{-1}(G_i d\sigma_i)\Sigma^{-1} \\ &+\mathrm{tr}S\Sigma^{-1}(G_j d\sigma_j)\Sigma^{-1}(G_i d\sigma_i)\Sigma^{-1} \ . \end{aligned} \qquad (4.8.2.6)$$

Applying a special case of Theorem 3.8.1 to (4.8.2.6), we obtain

$$\frac{\partial^2 g}{\partial \hat{\sigma}_j \partial \hat{\sigma}_i} = -\mathrm{tr}\hat{\Sigma}^{-1} G_j \hat{\Sigma}^{-1} G_i + 2\mathrm{tr}S\hat{\Sigma}^{-1} G_j \hat{\Sigma}^{-1} G_i \hat{\Sigma}^{-1} \qquad (4.8.2.7)$$

$$i,j = 0,\ 1,\ \ldots,\ m.$$

These second derivatives may be arranged in the form of a symmetric matrix, which yields the required Hessian matrix.

Likelihood equations and the Hessian matrix for the linear structure

$$\Sigma^{-1} = \sum_{i=0}^{m} \psi_i H_i \ ,$$

where H_i are known matrices and ψ_i are unknown parameters, may be similarly obtained.

Consider

$$h(\psi_0, \psi_1, \ldots, \psi_m) = \log|\Sigma^{-1}| - trS\Sigma^{-1} , \qquad (4.8.2.8)$$

which is to be maximized with respect to $\psi_0, \psi_1, \ldots, \psi_m$.

Differentiating (4.8.2.8) with respect to ψ_i, i = 0, 1, ..., m, gives

$$dh = tr\Sigma^T d(\Sigma^{-1})^T - trSd\Sigma^{-1} \qquad (4.8.2.9)$$

$$= \sum_{i=0}^{m} tr\Sigma H_i d\psi_i - \sum_{i=0}^{m} trSH_i d\psi_i \ . \qquad (4.8.2.10)$$

From (4.8.2.10), we obtain $\frac{\partial h}{\partial \psi_i}$ which, on setting equal to zero, gives

$$tr\hat{\Sigma}H_i = trSH_i \ , \quad i = 0, 1, \ldots, m; \qquad (4.8.2.11)$$

where $\hat{\Sigma} = \sum_{i=0}^{m} \hat{\psi}_i H_i$; [see Anderson (1968, eqn.(39))]. He does not provide any mathematical detail for obtaining (4.8.2.11).

To obtain a typical element of the Hessian matrix, we have

$$d^2_{\psi_j,\psi_i} h = -tr\Sigma^T H_j(d\psi_j)\Sigma^T H_i(d\psi_i) \qquad (4.8.2.12)$$

which implies

$$\frac{\partial^2 h}{\partial\hat{\psi}_j \partial\hat{\psi}_i} = -tr\left(\sum_{i=0}^{m} \hat{\psi}_i H_i\right)^{-1} H_j \left(\sum_{i=0}^{m} \hat{\psi}_i H_i\right)^{-1} H_i \ , \qquad (4.8.2.13)$$

where i,j = 0, 1, ..., m.

4.8.3 A Simpler Proof for a Result of Gleser & Olkin (1966)

We note that for any A:a×b,

$$\sum_{i=1}^{a}\sum_{j=1}^{b} a_{ij}^2 = \mathrm{tr}AA^T. \qquad (4.8.3.1)$$

We apply (4.8.3.1) in proving the following

Lemma 4.8.3.1 [Gleser & Olkin (1966, pp. 68-69)] If $L = (\ell_{ij})$:a×b and C:c×a, then

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\frac{\partial \log|W|}{\partial \ell_{ij}}\right)^2 = 4\mathrm{tr}LL^T(C^TW^{-1}C)^2, \qquad (4.8.3.2)$$

where $W = I+CLL^TC^T$.

Proof: Taking differentials, we get

$$\begin{aligned} d\log|W| &= \mathrm{tr}W^{-T}dW^T \\ &= \mathrm{tr}W^{-1}\{C(dL)L^TC^T+CL(dL^T)C^T\}. \end{aligned} \qquad (4.8.3.3)$$

Using Theorem 3.4.3 in (4.8.3.3), we obtain

$$\frac{\partial \log|W|}{\partial L} = 2C^TW^{-1}CL. \qquad (4.8.3.4)$$

An application of (4.8.3.1) in (4.8.3.4) gives

$$\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\frac{\partial \log|W|}{\partial \ell_{ij}}\right)^2 = 4\mathrm{tr}C^TW^{-1}CLL^TC^TW^{-1}C = 4\mathrm{tr}LL^T(C^TW^{-1}C)^2$$

which proves (4.8.3.2).

This lemma has been used by Gleser & Olkin (1966, pp. 67-68) in deriving the asymptotic non-central distribution of the likelihood-ratio statistic for testing the equality of vector parameters of a k-sample regression model with covariance. Gleser & Olkin (1966) provide the proof of this lemma by using elementwise differentiation of a scalar function of a matrix, which is quite lengthy and involved.

4.8.4 An Application of Matrix Derivatives to a Dynamic Econometric System

In this subsection we obtain certain matrix derivatives of the asymptotic covariance matrix S of the endogeneous vector variable for the following three forms of a dynamic econometric model considered by Conlisk (1969):

reduced form $$S = ASA^T+V \tag{4.8.4.1}$$

asymptotic form $$S = \sum_{\alpha=0}^{\infty} A^\alpha V(A^T)^\alpha \tag{4.8.4.2}$$

structural form $$S = C^{-1}DSD^TC^{-T}+C^{-1}UC^{-T} \ . \tag{4.8.4.3}$$

We observe that S is a function of A and V in the reduced and the asymptotic forms, and it is a function of C, D and U in the structural form.

First we show that the asymptotic and reduced forms are equivalent:

$$S = \sum_{\alpha=0}^{\infty} A^\alpha V(A^T)^\alpha \Longrightarrow S_r = \sum_{\alpha=0}^{\infty} (A\otimes A)^\alpha V_r,$$

using Theorem 2.6.3.1,

$$= (I-A\otimes A)^{-1}V_r, \text{ using (A.1.50).}$$

Hence

$$S_r-(A\otimes A)S_r = V_r \Longrightarrow S = ASA^T+V, \text{ which is the reduced form of S.}$$

To obtain the partial matrix derivatives of S with respect to A and V, it is more convenient to consider the reduced form (4.8.4.1). Let A be m×m matrix. Then differentiating S with respect to A and V, we obtain

$$dS = (dA)SA^T + A(dS)A^T + AS(dA^T) + dV,$$

which gives

$$dS_r = (I-A\otimes A)^{-1}[I\otimes AS+(AS\otimes I)I_{(m)}]dA_r + (I-A\otimes A)^{-1}dV_r. \quad (4.8.4.4)$$

From (4.8.4.4) we identify the required partial matrix derivatives as:

$$\frac{\partial S_r}{\partial A_r} = [I\otimes SA^T + I_{(m)}(SA^T\otimes I)](I-A^T\otimes A^T)^{-1}$$

$$= I\otimes SA^T(I+I_{(m)})(I-A^T\otimes A^T)^{-1}, \quad (4.8.4.5)$$

using (A.1.14).

$$\frac{\partial S_r}{\partial V_r} = (I-A^T\otimes A^T)^{-1}. \quad (4.8.4.6)$$

Differentiating the structural form of the asymptotic covariance matrix with respect to C, D and U, we get

$$dS = -C^{-1}(dC)C^{-1}DSD^TC^{-T} + C^{-1}(dD)SD^TC^{-T} + C^{-1}D(dS)D^TC^{-T}$$

$$+C^{-1}DS(dD^T)C^{-T} - C^{-1}DSD^TC^{-T}(dC^T)C^{-T} - C^{-1}(dC)C^{-1}UC^{-T}$$

$$+C^{-1}(dU)C^{-T} - C^{-1}UC^{-T}(dC^T)C^{-T}. \quad (4.8.4.7)$$

Here we note that $C{:}m\times m$, $D{:}m\times m$ and ${I_{(m)}}^T = I_{(m)}$.

Applying Theorem 2.6.3.1 in (4.8.4.7), we obtain, after some simplification,

$$dS_r = (I-C^{-1}D\otimes C^{-1}D)^{-1}[-\{C^{-1}\otimes C^{-1}DSD^TC^{-T} + (C^{-1}DSD^TC^{-T}\otimes C^{-1})I_{(m)}$$

$$+C^{-1}\otimes C^{-1}UC^{-T} + (C^{-1}UC^{-T}\otimes C^{-1})I_{(m)}\}dC_r + \{C^{-1}\otimes C^{-1}DS$$

$$+(C^{-1}DS\otimes C^{-1})I_{(m)}\}dD_r + (C^{-1}\otimes C^{-1})dU_r]$$

$$= (C\otimes C - D\otimes D)^{-1}[-\{(I+I_{(m)})(I\otimes DSD^TC^{-T} + I\otimes UC^{-T})\}dC_r$$

$$+\{(I+I_{(m)})(I\otimes DS)\}dD_r+dU_r]. \quad (4.8.4.8)$$

Hence the matrix derivatives of S with respect to the structural parameters C, D and U are:

$$\frac{\partial S_r}{\partial C_r} = -[\{I\otimes(C^{-1}DSD^T+C^{-1}U)\}(I+I_{(m)})](C^T\otimes C^T-D^T\otimes D^T)^{-1}, \quad (4.8.4.9)$$

$$\frac{\partial S_r}{\partial D_r} = \{(I\otimes SD^T)(I+I_{(m)})\}(C^T\otimes C^T-D^T\otimes D^T)^{-1}, \quad (4.8.4.10)$$

and

$$\frac{\partial S_r}{\partial U_r} = (C^T\otimes C^T-D^T\otimes D^T)^{-1}. \quad (4.8.4.11)$$

For simplifying the above partial derivatives we have used some properties of the Kronecker product, especially (A.1.14), and the result

$$(I-C^{-1}D\otimes C^{-1}D)^{-1} = (C\otimes C-D\otimes D)^{-1}(C\otimes C).$$

Conlisk (1969) obtained expressions for the partial derivatives $\frac{\partial L(S)}{\partial a_{ij}}$, $\frac{\partial L(S)}{\partial L(V)}$, $\frac{\partial L(S)}{\partial c_{ij}}$, $\frac{\partial L(S)}{\partial d_{ij}}$ and $\frac{\partial L(S)}{\partial L(U)}$, where $V = V^T$, $U = U^T$ and $L(S) = S_c$ in the notation of Tracy & Dwyer (1969). He derived these equations by expressing the asymptotic and the structural form of the covariance matrix S as a function of L(V) and L(U) respectively. His method is interesting only for linear functions of the parameter matrices.

We have followed a more general approach which is directly applicable to the reduced model and the structural model. We have not treated the asymptotic form, since it is equivalent to the reduced form and is much more complicated to handle than the reduced form. It may be observed that the expressions obtained by Conlisk (1969) are less general and more complicated than the equations (4.8.4.5), (4.8.4.6) and (4.8.4.9)-(4.8.4.11) presented here.

Neudecker (1969b) attempts to provide an expression for $\frac{\partial L(S)}{\partial L(A)}$ as an extension of $\frac{\partial L(S)}{\partial a_{ij}}$, given in Conlisk (1969, p. 279). His expression for $\frac{\partial L(S)}{\partial L(A)}$ involves a partitioned matrix of m row blocks which may be simplified by an application of $I_{(k)}$ given in the paper of Tracy & Dwyer (1969), [see Section 2.4 for various properties of $I_{(k)}$], and property (A.1.14) of the Kronecker matrix product.

CHAPTER V

POSSIBILITIES FOR FURTHER RESEARCH

Based on our discussion in Chapters II-IV, we may pursue further investigations in the theory of matrix differentiation and its applications.

In Theorem 2.4.1 some properties of the auxiliary matrices $I_{(k)}$ are stated. Results in matrix differentiation may be simplified if it is possible to express $I_{(m)}+I_{(n)}$ as some other kind of well-defined auxiliary matrix. Also, it is well-known that an orthogonal matrix P is proper if $|P| = 1$ and it is improper if $|P| = -1$ [see Murdoch (1971, p. 174)]. It is seen that for certain mn, $|I_{(m)}|$ and $|I_{(n)}|$ are either both proper or both improper. Using this fact one may characterize $I_{(m)}$ and $I_{(n)}$, which leads to investigating an integer function $m = f(m,n)$.

For the partitioned situation we observed that ${}_{\{m\}}I$ and $I_{\{n\}}$, introduced in Section 2.8, behave as $I_{(m)}$ and $I_{(n)}$ do in the non-partitioned situation. Hence a study of the nature of ${}_{\{m\}}I+I_{\{n\}}$, $|{}_{\{m\}}I|$ and $|I_{\{n\}}|$ may be useful in those situations where partitioned matrix differentiation is applicable.

In Section 4.8.1 we introduced auxiliary operators $I_{<q_i,i=1,2,\ldots,r;n>}$ and $I_{<n_i,i=1,2,\ldots,r;q>}$ and studied their properties and applications. It is worthwhile to express $I_{<q_i,i=1,2,..,r;n>}+I_{<n_i,i=1,2,\ldots,r;q>}$ as another auxiliary matrix which is a symmetric partitioned matrix. The resulting auxiliary matrix may have some additional properties and may be useful in simplifying matrix differentiation under very special situations where the matrix products $\odot$ and ⓔ are involved.

In Appendix A.1 we have listed various important properties of the Kronecker matrix product $\otimes$. Some of these properties are also true for our partitioned Kronecker product ⓟ, defined in Section 2.7. It may be desirable to verify additional properties of the matrix product ⓟ which are analogous to (A.1.27)-(A.1.51).

Madansky & Olkin (1968) have developed an alternative procedure for obtaining an asymptotic confidence region for a constraint vector parameter $\underline{h}(\underline{\theta})$, where $\underline{h}(\underline{\theta})$ is a vector function of a parameter vector $\underline{\theta}$. For various cases they have used elementwise differentiation to obtain approximate confidence intervals. One may apply the results of Chapter II and Chapter III to simplify the proofs of certain results in Madansky & Olkin (1968) and also to extend their procedure to set up an asymptotic confidence region for the constraint matrix parameter $H(\Theta)$, where $H(\Theta)$ is a matrix function of a parameter matrix Θ.

Lockhart (1967, pp. 268-271) has obtained typical elements of the asymptotic covariance matrix of the parameter estimators in the maximum likelihood factor analysis. The present matrix derivative methods may be applied in obtaining the asymptotic covariance matrix of the maximum likelihood estimators for the likelihood model (4.6.3.1). In particular, these methods will simplify the derivation presented by Lockhart (1967).

APPENDIX

Some important matrix results, many of which are used in this dissertation, are presented below.

A.1 The Kronecker Product of Matrices

Definition A.1.1 Let $A = (a_{ij})$:m×n matrix and B:p×q matrix; then the Kronecker product of A and B, written $A \otimes B$, is defined as the mp×nq matrix

$$A \otimes B = (a_{ij}B).$$

The proofs of the following results involving the Kronecker product $\otimes$ are straightforward:

$A \otimes B$, $B \otimes A$ exist for any A, B (A.1.1)

$A \otimes B \neq B \otimes A$ in general (A.1.2)

$A \otimes 0 = 0 \otimes A = 0$ (A.1.3)

$(\alpha A) \otimes B = \alpha(A \otimes B) = (A \otimes \alpha B)$, α is a scalar (A.1.4)

$A \otimes (B \otimes C) = (A \otimes B) \otimes C$ (A.1.5)

$(A \otimes B)(C \otimes D) = AC \otimes BD$ if AC, BD exist (A.1.6)

$(A \otimes B)^T = A^T \otimes B^T$ (A.1.7)

If A+B and C+D exist, then

$(A+B) \otimes (C+D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D$ (A.1.8)

$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, if the inverses exist (A.1.9)

$A^- \otimes B^-$ is a g-inverse of $(A \otimes B)$ for any A and B (A.1.10)

If A:m×m, B:n×n; then

$(I_m \otimes B)(A \otimes I_n) = (A \otimes I_n)(I_m \otimes B) = A \otimes B$ (A.1.11)

$I_m \otimes I_n = I_n \otimes I_m = I_{mn}$ (A.1.12)

$tr(A \otimes B) = (trA)(trB)$ (A.1.13)

If A:m×n, B:p×q matrices, then

(i) $I_{(p)}(A^T \otimes B) = (B \otimes A^T)I_{(q)}$ (A.1.14)

(ii) $I_{(p)}(A^T \otimes B)I_{(m)} = B \otimes A^T$ (A.1.15)

If X:m×n, Y:p×q, A:q×m, B:n×p matrices, then

(i) $trAXBY = (X_r)^T(A^T \otimes B)I_{(q)}Y_r$ (A.1.16)

(ii) $trAXBY = (Y_c)^T(A \otimes B^T)I_{(m)}X_c$ (A.1.17)

$Y = AXB \Longrightarrow Y_r = (A \otimes B^T)X_r$ (A.1.18)

$Y = AXB \Longrightarrow Y_c = (B^T \otimes A)X_c$ (A.1.19)

If A and B are symmetric matrices, then so is A⊗B. (A.1.20)

If A and B are skew-symmetric matrices, then A⊗B is a symmetric matrix. (A.1.21)

If A, B are symmetric and C, D are skew-symmetric matrices, then

A⊗B⊗C is a skew-symmetric matrix (A.1.22)

A⊗C⊗D is a symmetric matrix. (A.1.23)

If A and B are orthogonal matrices, then so is A⊗B. (A.1.24)

If A and B are upper (lower) triangular (diagonal) matrices, then so is A⊗B. (A.1.25)

If A is upper (lower) triangular and B is arbitrary, then A⊗B is a partitioned upper (lower) triangular matrix. (A.1.26)

$|A \otimes B| = |B \otimes A| = |A|^n |B|^m$, where A:m×m, B:n×n. (A.1.27)

$$(I-A\otimes B)^{-1} = \sum_{\alpha=0}^{\infty} (A\otimes B)^{\alpha} \text{ for square matrices A and B} \tag{A.1.28}$$

If $A{:}m\times m$ is idempotent, then so is $A\otimes A\otimes \ldots \otimes A$. (A.1.29)

If $A{:}m\times m$ is nilpotent, then so is $A\otimes A\otimes \ldots \otimes A$. (A.1.30)

If α_i, $i = 1, 2, \ldots, m$ are the latent roots of A, and if β_j, $j = 1, 2, \ldots, n$ are the latent roots of B; then

$\alpha_i\beta_j$ are the latent roots of both $A\otimes B$ and $B\otimes A$. (A.1.31)

If $\underline{x}_i$, $\underline{u}_i$ are the latent row vectors and latent column vectors of $A{:}m\times m$, and if $\underline{y}_j, \underline{w}_j$ are the latent row vectors and latent column vectors of $B{:}n\times n$, then the

latent row vectors of $A\otimes B$ are $\underline{x}_i\otimes y_j$ (A.1.32)

latent column vectors of $A\otimes B$ are $\underline{u}_i\otimes\underline{w}_j$ (A.1.33)

$\text{Rank}(A\otimes B) = (\text{Rank } A)(\text{Rank } B)$ (A.1.34)

If A and B are positive (positive semi-) definite matrices; then so is $A\otimes B$. (A.1.35)

If $A{:}r\times m$, $B{:}r\times n$ are given matrices such that $\text{Rank}(A\otimes B \vdots B\otimes A) = r^2$, then

$\text{Rank } A = \text{Rank } B = r$. (A.1.36)

Let $A_i{:}m_i\times m$, $B_i^T = (A_1^T,\ldots,A_{i-1}^T,A_{i+1}^T,\ldots,A_k^T)$ for $i = 1, 2, \ldots, k$, be given matrices such that

$\text{Rank } B_i = m$ for all i.

If

$$Q = A^T \circledcirc B^T \quad \text{where} \quad A = \begin{bmatrix} A_1 \\ \vdots \\ A_k \end{bmatrix}, \; B = \begin{bmatrix} B_1 \\ \vdots \\ B_k \end{bmatrix}, \text{ then}$$

$\text{Rank } Q = m^2$. (A.1.37)

(A.1.36) and (A.1.37) are slight modifications of properties (ii) and (iii) given by Khatri (1971).

If $A:m\times n$, $B:p\times q$, $P:r\times m$ and $Q:s\times p$ matrices, then

$$(P\otimes Q)(A \circledcirc B) = PA \circledcirc QB \qquad (A.1.38)$$

For partitioned matrices $A = (A_1,\ldots,A_k)$, $B = (B_1,\ldots,B_k)$; if $\text{Rank } A = \sum_{i=1}^{k} \text{Rank } A_i$ or $\text{Rank } B = \sum_{i=1}^{k} \text{Rank } B_i$, then

$$\text{Rank}(A \circledcirc B) = \sum_{i=1}^{k} \text{Rank}(A_i\otimes B_i) \qquad (A.1.39)$$

Proofs of (A.1.36) -(A.1.39) are very difficult and are given by Khatri (1971). Some of the other results given above are scattered in the following references: Graybill (1969) and Neudecker (1968, 1969a, 1969b). Besides these, there are other results concerning the Kronecker product $\otimes$ which are very easily derived from the above properties and are listed in Graybill (1969).

$$\text{Log}[I_m\otimes(I-A)] = I_m\otimes\text{Log}(I-A) \qquad (A.1.40)$$

$$\text{Log}[(I-B)\otimes I_n] = \text{Log}(I-B)\otimes I_n \qquad (A.1.41)$$

$$e^{I_m\otimes A} = I_m\otimes e^{A} \qquad (A.1.42)$$

$$e^{B\otimes I_n} = e^{B}\otimes I_n , \qquad (A.1.43)$$

where A and B are square matrices and m, n are arbitrary integers.

If $A:m\times m$, $B:n\times n$, then

$$e^{(I_n\otimes A)+(B\otimes I_m)} = e^{(I_n\otimes A)}e^{(B\otimes I_m)}. \qquad (A.1.44)$$

Results (A.1.45) and (A.1.46) are available in MacDuffee

(1946, pp. 83-84). These are the following:

If $A{:}n\times n$, $B{:}n\times n$ are symmetric matrices, then

$$\mathrm{sgn}(A\otimes B) = \mathrm{sgn}(A)\mathrm{sgn}(B) , \tag{A.1.45}$$

where sgn(A) denotes the signature of $A{:}n\times n$.

If α_i, $i = 1, 2, \ldots, m$ and β_j, $j = 1, 2, \ldots, n$ are latent roots of $A{:}m\times m$ and $B{:}n\times n$ respectively, then

$$\phi(\alpha_i,\beta_j) \text{ are the latent roots of both } \Phi(A;B) \text{ and } \Phi(B;A); \tag{A.1.46}$$

where

$$\Phi(A;B) = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}(A^i\otimes B^j)$$

and c_{ij} are real coefficients, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$.

If the matrix norm

$$\|A\| = \sqrt{\mathrm{tr}AA^T}$$

then

$$\|A\otimes B\| = \|A\| \; \|B\| . \tag{A.1.47}$$

$$|I_m\otimes(A-B)+J\otimes B| = |A-B|^{m-1}|A+(m-1)B|, \tag{A.1.48}$$

where $A{:}n\times n$, $B{:}n\times n$, and J is an $m\times m$ matrix with all its entries unity. This property is taken from Duthie (1971, p.95).

If $A = A^T$ is a non-singular matrix with all its entries rational numbers, then for any prime number p,

$$C_p(I_m\otimes A) = \{(-1,-1)_p\}^{m-1}\{C_p(A)\}^m\{(|A|,-1)_p\}^{m(m-1)/2} \tag{A.1.49}$$

where $C_p(X)$ is the Hasse-Minkowski p-invariant of A.

A detailed description of property (A.1.49) is presented by Bose & Connor (1952, pp. 376-377).

Let $P_A = AA^-$ be a projection operator, where

$P_A A = A$. (This follows from the definition of g-inverse of A).

Then for any two matrices A and B,

$$P_{A\otimes B} = P_A \otimes P_B \ , \qquad (A.1.50)$$

see also Rao & Mitra (1971, p. 119).

If, for any matrix $B: m^2 \times n^2$, there exists a matrix $A: m \times n$ such that

$$B = A \otimes A, \qquad (A.1.51)$$

then A is the square root of B with respect to the Kronecker matrix product.

A.2 Trace Properties of a Square Matrix

Definition A.2.1 Let $A = (a_{ij}): m \times m$ be a square matrix. Then the sum of the diagonal entries of A is known as the trace of A, which we denote as trA. Symbolically,

$$trA = \sum_{i=1}^{m} a_{ii}.$$

An excellant treatment of traces of square matrices is given by Graybill (1969). Below, we provide some of the trace properties:

$$trA = trA^T \qquad (A.2.1)$$

$$trAB = trBA, \ A: m \times n, \ B: n \times m \qquad (A.2.2)$$

$$trA^T B = trAB^T, \ A: m \times n, \ B: m \times n \qquad (A.2.3)$$

If $A: m \times m$ matrix whose latent roots are $\lambda_1, \lambda_2, \ldots, \lambda_m$, and k is a positive integer, then

$$tr(A^k) = \sum_{i=1}^{m} \lambda_i^k \ . \qquad (A.2.4)$$

$$trA = trP^{-1}AP, \ A:m\times m, \text{ and } P:m\times m \text{ is a non-singular matrix.} \qquad (A.2.5)$$

$$tr(\alpha A \pm \beta B) = \alpha trA \pm \beta trB, \ A:m\times m, \ B:m\times m \qquad (A.2.6)$$

$$tr(A^q B^q) = tr(B^q A^q), \ A:m\times m, \ B:m\times m \qquad (A.2.7)$$

For any A:m×n, we have

$$trAA^T = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \ . \qquad (A.2.8)$$

REFERENCES

Aitken, A. C. (1953). A note on trace-differentiation and the Ω-operator. Proc. Edinburgh Math. Soc. Ser. 2, 10 1-4.

Allen, D. M. (1967). Multivariate Analysis of Non-linear Models. Institute of Statistics Mimeo Series No. 558. University of North Carolina.

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

Anderson, T. W. (1968). Statistical inference for covariances with linear structures. In Multivariate Analysis-II (P. R. Krishnaiah, Ed.), Academic Press, New York, 55-66.

Anderson, T. W. & Rubin, H. (1956). Statistical inference in factor analysis. Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability, Vol. V 111-150.

Bock, R. D. & Bargmann, R. E. (1966). Analysis of covariance structures. Psychometrika 31 507-534.

Bose, R. C. & Connor, W. S. (1952). Combinatorial properties of group divisible incomplete block designs. Ann. Math. Statist. 23 367-383.

Capelli, A. (1887). Uber die Zurückführung des Cayleyschen operatoren Ω auf gewohnliche polaroperatoren. Mathematische Annalen 29 331-338.

Cayley, A. (1845). On the theory of linear transformations. Cambridge Mathematical Journal 4 193-209.

Cayley, A. (1846). On linear transformations. Cambridge and Dublin Mathematical Journal 1 104-122.

Conlisk, J. (1969). The equilibrium covariance matrix of dynamic econometric models. J. Amer. Statist. Assoc. 64 277-279.

Coy, J. W. (1955). A Differential Calculus for Functions of Matrices. Doctoral Dissertation, University of Michigan.

Deemer, W. L. & Olkin, I. (1951). The jacobian of certain matrix transformations useful in multivariate analysis. Based on the lectures of P. L. Hsu at the University of North Carolina, 1947. Biometrika 38 345-367.

Dirac, P. A. M. (1926). Quantum Algebra. Proc. Cambridge Phil. Soc. 23 412-418.

Duthie, A. I. (1971). Quintic Designs. Doctoral Dissertation, University of Windsor.

Dwyer, P. S. (1958). Generalizations of a Gaussian theorem. Ann. Math. Statist. 29 106-117.

Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. J. Amer. Statist. Assoc. 62 607-625.

Dwyer, P. S. (1970). The Concept of Matrix Element of a Matrix with Application to Matrix Derivatives. Unpublished Work.

Dwyer, P. S. & MacPhail, M. S. (1948). Symbolic matrix Derivatives. Ann. Math. Statist. 19 517-534.

Eisenpress, H. & Greenstadt, J. (1966). The estimation of non-linear econometric systems. Econometrica 34 851-861.

Ferrar, W. L. (1951). Finite Matrices. Oxford University Press, Oxford.

Fisk, P. R. (1967). Stochastically Dependent Equations. Charles Griffin and Company, London.

Fletcher, R. & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. Computer Journal 6 163-168.

Fleming, W. H. (1965). Functions of Several Variables. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts.

Fraser, R. A., Duncan, W. J. & Collar, A. R. (1936). Elementary Matrices and some Applications to Dynamics and Differential Equations. Cambridge University Press, Cambridge.

Garding, L. (1948). Extension of a formula by Cayley to symmetric determinants. Proc. Edinburgh Math. Soc. Ser. 2, 8 73-75.

Gebhardt, F. (1971). Maximum likelihood solution to factor analysis when some factors are completely specified. Psychometrika 36 155-163.

Gillespie, R. P. (1951). Partial Differentiation. Oliver and Boyd, Edinburgh.

Gleser, L. & Olkin, I. (1966). A k-sample regression model with covariance. In Multivariate Analysis (P. R. Krishnaiah, Ed.), Academic Press, New York, 59-72.

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

Goldberger, A. S., Nagar, A. L. & Odeh, H. S. (1961). The covariance matrices of reduced form coefficients and of forecasts for structural econometric model. Econometrica 29 556-573.

Gordon, F. S. (1967). Characterizations of Univariate and Multivariate Distributions using Regression Properties. Doctoral Dissertation, McGill University.

Gordon, F. S. & Mathai, A. M. (1972). Characterizations of the multivariate normal distribution using regression properties. Ann. Math. Statist. 43 205-229.

Graybill, F. A. (1969). Introduction to Matrices with Applications in Statistics. Wadsworth Publishing Company, Inc., Belmount, California.

Harman, H. H. (1967). Modern Factor Analysis. Second Edition (Revised). University of Chicago Press, Chicago.

Hoffman, K. & Kunze, R. (1971). Linear Algebra. Second Edition. Prentice-Hall, Englewood Cliffs, N. J.

Hsu, P. L. (1953). On symmetric, orthogonal and skew-symmetric matrices. Proc. Edinburgh Math. Soc. Ser. 2, 10 37-44.

Hua, L. K. (1958). Harmonic Analysis of Functions of Several Complex Variables in the Classical Domains. Original Chinese text published by Science Press, Peking, Translations of Mathematical Monographs, Vol. 6, Amer. Math. Soc., Providence, Rhode Island, 1963.

Jack, H. (1964-65). Jacobian of transformations involving orthogonal matrices. Proc. Roy. Soc. Edinburgh Section A, 67 81-103.

Johnston, J. (1963). Econometric Methods. McGraw-Hill Book Company, Inc., New York.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. Psychometrika 31 165-178.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. Psychometrika 32 443-482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34 183-202.

Jöreskog, K. G. (1970a). A general method for analysis of covariance structures. Biometrika 57 239-251.

Jöreskog, K. G. (1970b). Estimation and testing of simplex models. British Journal of Mathematical and Statistical Psychology 23 121-145.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika 36 109-133.

Jöreskog, K. G., Gruvaeus, G. T. & van Thillo, M. (1970). ACOVS-A General Computer Program for Analysis of Covariance Structures. Research Bulletin 70-15. Educational Testing Service, Princeton, New Jersey.

Jöreskog, K. G. & Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. British Journal of Mathematical and Statistical Psychology 21 85-96.

Khatri, C. G. (1965). Classical statistical analysis based on a certain multivariate complex Gaussian distribution. Ann. Math. Statist. 36 98-114.

Khatri, C. G. (1968). Some Jacobian Transformations and their Applications in Multivariate Distribution Theory. Unpublished Work.

Khatri, C. G. (1971). On Characterization of gamma and multivariate normal distributions by solving some functional equations in vector variables. J. Multivariate Analysis 1 70-89.

Khatri, C. G. & Rao, C. R. (1968). Solutions to some functional equations and their applications to characterization of probability distributions. Sankhyá Ser. A, 30 167-180.

Kleinbaum, D. G. (1970). Estimation and Hypothesis Testing for Generalized Multivariate Linear Models. Doctoral Dissertation, University of North Carolina.

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

Lawley, D. N. (1967). Some new results in maximum likelihood factor analysis. Proc. Roy. Soc. Edinburgh Section A, 67 256-264.

Lawley, D. N. & Maxwell, A. E. (1963). Factor Analysis as a Statistical Method. Butterworth, London.

Lockhart, R. S. (1967). Asymptotic sampling variances for factor analytic models identified by specified zero parameters. Psychometrika 32 265-277.

MacDuffee, C. C. (1946). The Theory of Matrices. Chelsea Publishing Company, New York.

Madansky, A. & Olkin, I. (1968). Approximate confidence regions for constraint parameters. In Multivariate Analysis-II (P. R. Krishnaiah, Ed.), Academic Press, New York, 261-286.

Michal, A. D. (1947). Matrix and Tensor Calculus. John Wiley and Sons, Inc., New York.

Morrison, D. F. (1967). Multivariate Statistical Methods. McGraw-Hill Book Company, Inc., New York.

Mulaik, S. A. (1971). A note on some equations of confirmatory factor analysis. Psychometrika 36 63-70.

Murdoch, D. C. (1971). Linear Algebra for Undergraduates. John Wiley and Sons, Inc., New York.

Neudecker, H. (1967). On matrix procedures for optimizing differentiable scalar functions of matrices. Statistica Neerlandica 21 101-107.

Neudecker, H. (1968). The Kronecker matrix product and some of its applications in econometrics. Statistica Neerlandica 22 69-82.

Neudecker, H, (1969a). A note on Kronecker matrix products and matrix equation systems. Siam J. Appl. Math. 17 603-606.

Neudecker, H. (1969b). Some theorems on matrix differentiation with special reference to Kronecker matrix products. J. Amer. Statist. Assoc. 64 953-963.

Olkin, I. (1951). On Distribution Problems in Multivariate Analysis. Institute of Statistics Mimeograph Series, Report No. 8, University of North Carolina.

Olkin, I. (1953). Note on the Jacobian of certain matrix transformations useful in multivariate analysis. Biometrika 40 43-46.

Olkin, I. & Roy, S. N. (1954). On multivariate distribution theory. Ann. Math. Statist. 25 329-339.

Olkin, I. & Sampson, A. R. (1969). Jacobians of Matrix Transformations and Functional Equations. Technical Report No. 40, Department of Statistics, Stanford University.

Potthoff, R. F. & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful specially for growth problems. Biometrika 51 313-326.

Powers, D. L. (1971). On the differentials of certain matrix functions. Canadian J. Math. 23 282-286.

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

Rao, C. R. & Mitra, S. K. (1971). Generalized Inverse of Matrices and its Applications. John Wiley and Sons, Inc., New York.

Rinehart, R. F. (1966a). The differential of a primary matrix function. Rend. Circ. Mat. Palermo 15 209-215.

Rinehart, R. F. (1966b). P and D in $P^{-1}XP = dg\{\lambda_1, \lambda_2, \ldots, \lambda_n\} = D$ as matrix functions of X. Canadian J. Math. 18 832-837.

Rothenberg, T. J. & Leenders, C. T. (1964). Efficient estimation of simultaneous equation systems. Econometrica 32 57-76.

Roy, S. N. (1957). Some Aspects of Multivariate Analysis. John Wiley and Sons, Inc., New York.

Scheffé, H. (1959). The Analysis of Variance. John Wiley and Sons, Inc., New York.

Singh, R. P. & Tracy, D. S. (1970). Applications of matrix derivatives to estimation problems in factor analysis. Contributed Paper for the 1970 Annual Meeting of the American Statistical Association.

Srivastava, J. N. (1966). On testing hypotheses regarding a class of covariance structures. Psychometrika 31 147-164.

Stroud, T. W. F. (1968). Comparing Conditional Distributions under Measurement Errors of Known Variances. Technical Report No. 28, Department of Statistics, Stanford University.

Stroud, T. W. F. (1971). On obtaining large sample tests from asymptotically normal estimators. Ann. Math. Statist. 42 1412-1424.

Tan, W. Y. (1968-69). Some results on multivariate regression analysis. Nanta Mathematica 3 54-71.

Tracy, D. S. & Dwyer, P. S. (1969). Multivariate maxima and minima with matrix derivatives. J. Amer. Statist. Assoc. 64 1576-1594

Tracy, D. S. & Singh, R. P. (1970a). Differentiation of functions of matrices with scalar relationships among elements. Contributed Paper at the 14-th Ontario Mathematical Meeting on February 7, 1970.

Tracy, D. S. & Singh, R. P. (1970b). Matrix differentiation after partitioning, and other results. Ann. Math. Statist. 41 1150.

Tracy, D. S. & Singh, R. P. (1971a). Matrix derivative approach to analysis of covariance structures. Proceedings of 38-th Session of International Statistical Institute 289-293.

Tracy, D. S. & Singh, R. P. (1971b). A new matrix product and its applications. Ann. Math. Statist. 42 1798.

Trawinski, I. M. & Bargmann, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. Ann. Math. Statist. 35 647-657.

Turnbull, H. W. (1927-29). On differentiating a matrix. Proc. Edinburgh Math. Soc. Ser. 2, 1 111-128.

Turnbull, H. W. (1930-31a). A matrix form of Taylor's theorem. Proc. Edinburgh Math. Soc. Ser. 2, 2 33-54.

Turnbull, H. W. (1930-31b). Matrix differentiation of the characteristic function. Proc. Edinburgh Math. Soc. Ser. 2, 2 256-264.

Turnbull, H. W. (1948). Symmetric determinants and the Cayley and Capelli operators. Proc. Edinburgh. Math. Soc. Ser. 2, 8 76-78.

Turnbull, H. W. (1949). Note upon the generalized Cayleyan operator. Canadian J. Math. 1 48-56.

Wallace, A. H. (1953). A note on the Capelli operators associated with a symmetric matrix. Proc. Edinburgh Math. Soc. Ser. 2, 9 7-12.

Wilks, S. S. (1962). Mathematical Statistics. John Wiley and Sons, Inc., New York.

Wrobleski, W. J. (1963). Extensions of the Dwyer-MacPhail Matrix Derivative Calculus with Applications to Estimation Problems Involving Errors-in-Variables and Errors-in-Equations. Technical Report, Office of Research Administration, University of Michigan.

VITA AUCTORIS

Birth	August 15, 1943	Varanasi, India
B. Sc. (Hons.)	1962	Banaras Hindu University
M. Sc.	1964	Banaras Hindu University
M. Sc.	1969	University of Windsor

At present Statistical Consultant, National Health Grant Project 606-21-72, School of Social Work, University of Windsor, Windsor, Ontario.