

Negative microfilm
received from university

64-913

WROBLESKI, William Joseph, 1929-
EXTENSIONS OF THE DWYER-MACPHAIL
MATRIX DERIVATIVE CALCULUS WITH APPLI-
CATIONS TO ESTIMATION PROBLEMS INVOLVING
ERRORS-IN-VARIABLES AND ERRORS-IN-
EQUATIONS.

The University of Michigan, Ph.D., 1963

Mathematics
University Microfilms, Inc., Ann Arbor, Michigan

EXTENSIONS OF THE DWYER-MACPHAIL MATRIX DERIVATIVE CALCULUS
WITH APPLICATIONS TO ESTIMATION PROBLEMS
INVOLVING ERRORS-IN-VARIABLES AND ERRORS-IN-EQUATIONS

by

William Joseph Wroblewski

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in The
University of Michigan
1963

Doctoral Committee:

Professor Paul S. Dwyer, Chairman
Professor Cecil C. Craig
Associate Professor Bernard A. Galler
Assistant Professor Bruce M. Hill
Professor W. Allen Spivey

ACKNOWLEDGMENTS

The author takes this opportunity to express his most sincere appreciation to Professor Paul S. Dwyer for his sustained encouragement and guidance which made this research possible. This influence was felt not only through his role as doctoral committee chairman, but it began years ago when, as a teacher, he first introduced the author to statistics with such unlimited enthusiasm, patience and insight that a self-sustaining force was provided for carrying the original motivations he inspired to a fuller maturity in the form of this thesis.

In addition the author also wishes to express a warm acknowledgment to the other members of his doctoral committee, each of whom took time from demanding schedules to read a draft of this manuscript and to provide helpful advice and corrections. The author, however, assumes the responsibility for any omissions, lack of complete mathematical rigor or profundity, incongruous expressions, or other oversights that may still appear in this work.

The final typing of this thesis was done by Miss Beatrice C. Finley who worked under the pressure of a serious time limitation. Her exceptional sense of proportion and her skills in arranging the many symbols and equations which appear here attest to her proficiency in technical typing, and they spared the author from having to become deeply involved in arranging the report's typographical format. For all of these reasons the author is most grateful to her. The author also recognizes his indebtedness to the National Science Foundation for providing him with partial financial support under Grant No. GP-12 (G-18876).

In the way of a final acknowledgment the author, were he able, would like to express in a few simple words his deepest heartfelt feelings for the loyalty, unselfish sacrifice and complete support given him over long and trying, almost desperate, years by his wife and family: Virginia, Ann Marie, Mary Ellen, Carol Jean and William Thomas. Even under the burden of a growing family, Virginia still found it possible to type almost an entire first draft of this manuscript and then, at the end, to help proofread the final document. The author finds it impossible, however, to write down his eternal gratitude to his wife and family for sustaining him over an extremely difficult period; and so, these feelings must remain as silent sentiments to be understood now and hereafter only by those who shared them so intimately.

TABLE OF CONTENTS

CHAPTER	Page
1. OUTLINE OF RESEARCH TOPICS AND STATEMENT OF RESULTS	1
1.1. Introduction	1
1.2. Matrix Derivatives	2
1.3. Estimation Problems Involving Errors-in-Variables and/or Errors-in-Equations	12
1.4. Problems for Future Research	17
2. EXTENSIONS OF THE DWYER-MACPHAIL MATRIX DERIVATIVE	20
2.1. Introduction	20
2.2. The Dwyer-Macphail Matrix Derivative	20
2.3. Illustration of Dwyer-Macphail Matrix Differentiation	29
2.4. Generalization of a Useful Duality Type Theorem in the Calculus of Dwyer-Macphail Matrix De- rivatives	32
2.5. Higher Order Dwyer-Macphail Matrix Derivatives	52
2.6. Series Expansions Featuring Dwyer-Macphail Matrix Derivatives	82
3. DWYER-MACPHAIL MATRIX DERIVATIVES AND FRECHET-GATEAUX MATRIX DIFFERENTIALS	103
3.1. Introduction	103
3.2. Basic Algebraic and Topological Concepts	107
3.3. Gateaux and Frechet Differentials of Abstract Functions	112
3.4. Abstract Polynomials and Higher Order Gateaux and Frechet Differentials	120
3.5. A Basic Relation Between Dwyer-Macphail Matrix Derivatives and Frechet-Gateaux Matrix Differ- entials	127
4. APPLICATIONS TO ESTIMATION PROBLEMS IN STATISTICS	149
4.1. Introduction	149
4.2. Stochastic Structures and EV or EVE Statistical Models	154
4.3. Generalized Least Squares Estimators for EV Model	167
4.4. Generalized Least Squares Estimators for EVE Model	184
4.5. Regularity Conditions	197

TABLE OF CONTENTS (Concluded)

CHAPTER	Page
5. ESTIMATION IN MODIFIED EV AND EVE STATISTICAL MODELS	218
5.1. Introduction	218
5.2. Maximum Likelihood Estimators for the EV Model	221
5.3. Maximum Likelihood Estimators for the EVE Model	227
5.4. Connections With Other Research on the Estimation Problem in the EV and EVE Models	231
5.5. Summary of Research	252
BIBLIOGRAPHY	259

CHAPTER 1

OUTLINE OF RESEARCH TOPICS AND STATEMENT OF RESULTS

1.1. INTRODUCTION

In this work, two research topics are studied, one in applied mathematics dealing with the notion of matrix derivatives and the other in statistics dealing with the application of matrix derivatives to maximization and minimization problems connected with estimating the unknown parameters of a linear structural relationship having errors-in-variables and/or errors-in-equations. Chapters 2 and 3 pertain directly to the topic of matrix derivatives while Chapters 4 and 5 are concerned with the application of these notions to estimation problems in statistics. Thus the reader who is interested in exploring techniques of matrix differentiation of matrix functions of a matrix variable may confine his attention to reading only Chapters 2 and 3, while the reader more interested in the statistical applications may go directly to Chapters 4 and 5 although in places he may find it necessary to refer to certain theorems stated in the preceding two chapters in order to follow each detail of the derivations given there. Accordingly, for convenience in exposition, we review separately in Sections 1.2 and 1.3 the topics of matrix derivatives and their applications to estimation problems in statistics involving errors-in-variables and/or errors-in-equations and summarize the new results that have been obtained in these areas in applied mathematics and statistics. In addition, in Section 1.4 we briefly list several

problems related to these topics as a basis for future research.

1.2. MATRIX DERIVATIVES

The notion of "derivative" of a matrix function of a matrix variable has been studied in applied mathematics by several authors; and we make no attempt here to provide a complete survey of these works, although in [23] a fairly extensive bibliography on this topic current to 1955 can be found. Rather we concentrate our attention on one such notion of matrix derivative which has proven especially relevant to applications in statistics. In many problems in statistics one is frequently faced with the task of maximizing or minimizing a function of a matrix variable or of expanding some complicated functional of a matrix variable in a meaningful series expansion in order to achieve a useful approximation. In these types of problems, therefore, there is found ample motivation for providing matrix differentiation techniques which enable one to obtain explicit matrix expressions for the derivative of a function of a matrix variable simultaneously with respect to each of the elements of the matrix variable.

Prior to the development of the matrix derivative suggested in [32] by Dwyer and Macphail, such extremal problems were usually handled by reinterpreting the matrix problem in terms of associated scalar problems, applying the ordinary differential calculus for real or complex functions of real or complex variables to the equivalent scalar problems and then, once again, reassembling the derived results into a matrix state-

ment, the details of these calculations generally being omitted from the mathematical presentation.

Thus, for example, when x is an $n \times 1$ column vector and A is an $n \times n$ -matrix, we find that Frazer, Duncan and Collar [35] give the result

$$\frac{\partial y}{\partial x} = 2Ax$$

as the derivative with respect to x of the quadratic form,

$$y = f(x) = x^T A x .$$

Here $\partial/\partial x$ may be interpreted as an operator having the same row and column orders as x whose components $\partial/\partial x_k$, $k = 1, 2, \dots, n$, are the ordinary scalar differentiation operators.

In order to provide a concise systematic foundation for using matrix derivatives in statistical applications so that matrix differentiation can be carried out in a straightforward manner without the necessity of reverting to and from equivalent scalar expressions and differentiation operations when examining extremal properties of matrix functions, Dwyer and Macphail in [32] developed a calculus of matrix differentiation for scalar functions of a matrix variable. Thus if $Y = F(X)$ is a $p \times q$ -matrix function of an $m \times n$ -matrix variable X , they define the derivative of the rs th element function $\langle Y \rangle_{rs} = f_{rs}(X)$ of $F(X) = [f_{rs}(X)]$ with respect to the matrix X simply as the $m \times n$ -matrix

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \begin{bmatrix} \frac{\partial f_{rs}(X)}{\partial x_{11}} & \frac{\partial f_{rs}(X)}{\partial x_{12}} & \dots & \frac{\partial f_{rs}(X)}{\partial x_{1n}} \\ \frac{\partial f_{rs}(X)}{\partial x_{21}} & \frac{\partial f_{rs}(X)}{\partial x_{22}} & \dots & \frac{\partial f_{rs}(X)}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_{rs}(X)}{\partial x_{m1}} & \frac{\partial f_{rs}(X)}{\partial x_{m2}} & \dots & \frac{\partial f_{rs}(X)}{\partial x_{mn}} \end{bmatrix}$$

where $\partial f_{rs}(X)/\partial x_{ij}$ is the ordinary partial derivative of $f_{rs}(X)$ interpreted as a function of several scalar variables x_{ij} which are the mn -elements of the matrix X .

In their original paper, Dwyer and Macphail concentrated on providing methods for deriving this type of matrix derivative for a wide class of functions given as sums of products of constant and variable matrices. They established three rules which involved certain special matrices and their transposes for obtaining this type of matrix derivative for such matrix functions. These rules could be used symbolically in the sense that one could interpret them as specifying when certain symbols were to be replaced by certain other symbols and how these symbols were to be arranged in order to obtain this type of matrix derivative.

It was the symbolic interpretation of these three rules prescribing how this type of matrix derivative could be written down from the ordinary matrix derivative of a related matrix function of a scalar variable

which led originally to these matrix derivatives being called "symbolic" matrix derivatives. It is this author's feeling, however, that the adjective "symbolic" in this context imputes a secondary or colored meaning to this basic mathematical notion of a matrix derivative of a scalar function of a matrix variable; and it is for this reason that this concept is called the "Dwyer-Macphail matrix derivative" throughout this presentation.

In addition to providing an almost mechanical basis for calculating such matrix derivatives, these rules also made it possible to derive the Dwyer-Macphail matrix derivative of a scalar function of a matrix variable without having to give explicit consideration to the representation of such a function as a scalar function of several scalar variables. In other words, if we desired to find the derivative of the function

$$y = \langle X^2 \rangle_{ij}, \quad \text{where } X \text{ is an } nxn\text{-matrix}$$

for example, then $\partial \langle X^2 \rangle_{ij} / \partial X$ could be obtained by applying these rules without having to write down explicitly the representation of y as a function of several scalar variables, namely

$$y = \sum_{k=1}^n x_{ik} x_{kj} = \langle X^2 \rangle_{ij},$$

and forming successively the partial derivatives $\partial y / \partial x_{rs}$ for $1 \leq r \leq n$ and $1 \leq s \leq n$.

The usefulness of this matrix derivative calculus is well illustrated in [31] where Dwyer uses the Dwyer-Macphail matrix derivative to

present for the first time self-contained derivations of several generalizations of a Gaussian theorem. In particular Dwyer is able to derive a general formula for simultaneous unbiased estimates of linear functions of parameters (when these parameters are subject to linear restrictions) and a formula for the variance matrix of these estimates. He proves, moreover, without arbitrarily having to introduce special matrix expressions which implicitly contain the desired results—but rather by the direct application of the Dwyer-Macphail matrix derivative methods—that these estimates have minimum variance among all unbiased linear estimates.

Although the calculus for functions of a matrix variable based on the notion of Dwyer-Macphail matrix derivatives has already been refined to the point where its application has led to important new results in statistics, nevertheless, there are some theoretical areas that remain to be investigated and these provide the main topics for this research in applied mathematics on matrix derivatives.

Our first result on matrix derivatives is related to an attempt to extend the three rules obtained by Dwyer and Macphail for deriving the Dwyer-Macphail matrix derivative of a function given as a sum of products of constant or variable matrices to a larger class of matrix functions of a matrix variable. We accomplish this by proving for the first time a duality type theorem between the ordinary (or J-type) matrix derivative and the Dwyer-Macphail (or K-type) matrix derivative of a general matrix function. In turn the three rules originally established by

Dwyer and Macphail have been rigorously established by an application of this more basic duality type theorem and certain properties of the Trace operator. Then this same duality type theorem is used to establish a new compound function differentiation theorem for matrix functions of a matrix variable which features the Dwyer-Macphail matrix derivatives of the functions defining the compound matrix function, and this matrix differentiation theorem for compound matrix functions plays an important role in our subsequent developments related to the estimation problem for stochastic structures representing linear structural relationships having errors-in-variables and errors-in-equations.

Having established these results for first order Dwyer-Macphail matrix derivatives, it is natural to inquire as to whether or not a satisfactory notion of higher order Dwyer-Macphail matrix derivative can be introduced which extends the theory developed for first order Dwyer-Macphail matrix derivatives while, at the same time, leading to identical theorems when first order matrix derivatives are considered. Following this line of inquiry, five types of higher order matrix derivatives have been defined and their properties studied.

To this author's knowledge, none of these five higher order matrix derivatives has been previously introduced or investigated. Our research has led to a generalized duality type theorem among these various types of higher order matrix derivatives which contains the duality theorem for first order matrix derivatives as a special case and which allows all of these five new types of higher order matrix derivatives to

be derived for many matrix functions by direct reapplication of the three simple rules which previously were known to be valid only for the first order matrix derivatives of the J or K types mentioned above.

In addition one of these five types of higher order matrix derivatives, namely higher order Dwyer-Macphail matrix derivatives, has provided a basis for developing a matrix function of a matrix variable in an infinite series expansion which explicitly features its matrix derivatives. An application of this extended calculus of Dwyer-Macphail matrix derivatives, for example, leads to the expansion

$$XAX^T = \bar{X}A\bar{X}^T + XA(X-\bar{X})^T + (X-\bar{X})AX^T + (X-\bar{X})A(X-\bar{X})^T$$

of the matrix function $Y = F(X) = XAX^T$ about the matrix $X = \bar{X}$ where A and X are $n \times n$ -matrices.

Thus the introduction of these new higher order Dwyer-Macphail matrix derivatives and their use in developing a matrix function of a matrix variable in a series expansion are two new extensions of the Dwyer-Macphail matrix derivative calculus which maintain the simplicity of the original development of the calculus of first order Dwyer-Macphail matrix derivatives.

These results are developed in Chapter 2; then in Chapter 3 we shift the orientation of our research on matrix derivatives quite sharply, this change of emphasis being motivated by the following question: What relationship exists between the Dwyer-Macphail matrix derivative and certain notions of matrix differentials which are adapted interpretations

to matrix spaces of differentials for functions from and to more general abstract spaces. In particular we seek the relationship between Dwyer-Macphail matrix derivatives and Gateaux or Frechet differentials [38,39, 40] of matrix functions of a matrix variable.

In several sections of Chapter 3 we summarize in their present form parts of the theory of Frechet and Gateaux differentials for functions from and to real or complex Banach spaces or Banach algebras, this summary dealing mainly with those parts of the theory which are especially significant to our work involving matrix derivatives and matrix differentials. We find, under certain regularity conditions, that the Frechet differential and the Gateaux differential of an abstract function co-exist and are co-equal; and in these circumstances we refer merely to the Frechet-Gateaux differential of the abstract function. The Frechet-Gateaux differential is in fact an abstract homogeneous polynomial of degree one (i.e., it is an abstract linear function). Similarly higher order Frechet-Gateaux differentials are abstract homogeneous polynomials of degree corresponding to their order (i.e., higher order Frechet-Gateaux differentials are abstract symmetric multilinear forms each of whose variables is assigned a common value).

In the case of a function from and to matrix Banach spaces or algebras over the real or complex fields, we can anticipate that its Frechet-Gateaux differential is functionally related to the ordinary partial derivatives of the element functions of the given matrix function ; and, therefore, since the Dwyer-Macphail matrix derivative is also de-

defined in terms of these same partial derivatives, that the two notions of Frechet-Gateaux matrix differentials and Dwyer-Macphail matrix derivatives are most certainly related to one another. But this is an oversimplification of the question which we raised earlier regarding the relationship between these two notions of matrix differentials and matrix derivatives.

Clearly, we can expect that these two matrix concepts are related to each other through the functional dependence of each of these two notions on the same ordinary partial derivatives of the element functions of the given matrix function, but the significant point is that the fundamental mathematical structure of the Frechet-Gateaux matrix differential as an abstract homogeneous matrix polynomial is directly, explicitly and uniquely determined in terms of the corresponding Dwyer-Macphail matrix derivatives of the matrix function. This cannot be said of other notions of matrix derivatives such as that studied by Rinehart [123]. In other words, although the Rinehart matrix derivative (again through the ordinary partial derivatives of the element functions of the given matrix function) is also functionally related to the Frechet-Gateaux matrix differential, nevertheless, this relationship cannot be expressed in terms of an abstract homogeneous matrix polynomial (i.e., by a symmetric, multilinear matrix form).

For our purposes the operational consequence of this observation is this: Many of the properties known to hold in the elementary calculus for derivatives and differentials of real functions of a real variable

can be extended to derivatives and differentials of matrix functions of a matrix variable by virtue of the special relationship that exists between Dwyer-Macphail matrix derivatives and Frechet-Gateaux matrix differentials. For example, in the elementary calculus the derivative of a function is zero if and only if its differential vanishes identically as a function of its increment, while for matrix functions of a matrix variable, the Dwyer-Macphail matrix derivatives vanish if and only if the Frechet-Gateaux matrix differential vanishes identically as a function of its (matrix) increment; and, therefore, necessary conditions for critical points of a scalar function of a matrix variable can be stated equivalently in terms of either matrix differentials of the Frechet-Gateaux type or in terms of matrix derivatives of the Dwyer-Macphail type.

These issues are clarified in the course of our treatment of matrix differentials and matrix derivatives found in Chapter 3, and it suffices to say at this point as a summary statement that new extensions of the Dwyer-Macphail matrix derivative have been obtained in the form of certain basic relationships which exist between them and the Gateaux and Frechet differentials of matrix functions of matrix variables.

Other notions of differentials for abstract functions have been proposed by various authors and their properties explored such as the notion of Lorch differential [94] for functions from and to commutative Banach algebras and Hausdorff differentials [57] for functions from and to associative linear algebras with a unit element over the real or com-

plex field. It appears, moreover, that the Rinehart matrix derivative may be related to the Hausdorff matrix differential just as the Dwyer-Macphail matrix derivative is related to the Frechet-Gateaux matrix differential.

We do not, however, pursue this or related investigations further in this dissertation. Rather we have mentioned the notions of Lorch differentials, Hausdorff differentials, and Rinehart matrix derivatives only to call the reader's attention to the fact that our discussions in no way exhaust all possible approaches to the study of matrix derivatives and matrix differentials. But this point notwithstanding, it does appear that the study of Dwyer-Macphail matrix derivatives and Frechet-Gateaux matrix differentials offers a rich selection of theory and techniques which are quite useful and adequate for solving many of the complicated types of optimization problems that occur in mathematical statistics.

1.3. ESTIMATION PROBLEMS INVOLVING ERRORS-IN-VARIABLES AND/OR ERRORS-IN-EQUATIONS

In Chapters 4 and 5 we turn our attention to applying the extended calculus of Dwyer-Macphail matrix derivatives to certain estimation problems in statistics which involve errors-in-variables and/or errors-in-equations. The basic statistical model with which we are concerned involves a linear relationship between sure and random variables in which random "errors-in-equations" or "shock" disturbances may be present and in which the sure and random variables of the structural relation are

non-observable, their values being masked by random "errors-in-variables."

When both variable and equation disturbances are present we say that we have an EVE (error-in-variable-and-equation) statistical model, namely a model of the form

$$\sum_{i=1}^p \alpha_i r_i(k) + \sum_{j=1}^q \beta_j S_j(k) = \xi(k)$$

where for $1 \leq k \leq n$ we observe

$$x_i(k) = r_i(k) + \delta_i(k) \quad \text{and} \quad y_j(k) = S_j(k) + \eta_j(k) .$$

Here the $r_i(k)$ are non-observable sure variables and the $S_j(k)$ are non-observable random variables, $\delta_i(k)$ and $\eta_j(k)$ being random "errors-in-variables" and $\xi(k)$ being an "error-in-equation;" and when the random elements $S_j(k)$ and $\xi(k)$ are zero with probability one, so that the linear structural relation is merely an implicit linear function,

$$\alpha_1 r_1(k) + \alpha_2 r_2(k) + \dots + \alpha_p r_p(k) = 0 ,$$

and for $1 \leq k \leq n$ we observe $x_i(k) = r_i(k) + \delta_i(k)$ instead of $r_i(k)$ directly, we say that we have an EV (error-in-variable) statistical model.

We apply the theory and techniques of matrix derivatives and matrix differentials as developed in Chapters 2 and 3 to the problem of estimating the unknown parameters of a linear structural relationship having errors-in-variables and/or errors-in-equations. In our development the variable and equation errors jointly are allowed to have general second order covariances rather than more simplified covariance structures

and associated, for example, with the assumption of independently distributed variable or equation errors.

In Chapter 4, adopting a generalized least squares estimation procedure whose weights turn out to depend on the unknown parameters being estimated, we find that such GLS (generalized least squares)-estimators of the unknown parameters of a linear structural relationship whose equation and variable errors are assumed to have a known but general covariance matrix, are given as critical points of a very complicated functional which is a compound function of vector and matrix functions on a certain vector space. Consequently, following a GLS-approach to this estimation problem, we are led to differentiating scalar, vector, and matrix functions of a vector variable, and thus we can expect the Dwyer-Macphail matrix derivative techniques to play a significant role in these investigations.

In the last chapter, we examine the estimation problem for modified EV and EVE statistical models in which additional assumptions are made about the probability distributions of the random elements of these two stochastic structures. One of the strongest additional specifications that we make is to assume that the equation or variable disturbances are jointly normally distributed. When the model errors are normally distributed, the maximum likelihood estimation procedure is readily applicable, of course, and we investigate the ML (maximum likelihood)-estimators of the unknown parameters of these modified EV and EVE statistical models in which additional assumptions are made about the variable and

equation errors.

As a summary of the results to be found in Chapters 4 and 5 on the estimation problem for EVE and EV statistical models, we briefly point out that maximum likelihood estimators or generalized least squares estimators have been derived for the first time for the unknown parameters of a linear structural relationship having both error-in-variables and errors-in-equations and without restrictions on the covariance matrices of the unknown errors, except of course, that the errors are a second order random process (i.e., that their variances and covariances are finite); and the maximum likelihood estimators of the structural parameters of a linear structural relationship are shown to be identical to the generalized least squares estimators of these parameters.

In addition, by means of the relationships existing between Dwyer-Macphail matrix derivatives and Frechet-Gateaux matrix differentials we have established not only the necessary conditions for critical points of the complicated functional associated with GLS-estimation of the unknown parameters of an EV or EVE statistical model but, as important, we also have established a sufficient condition for a critical point of this functional to be a local minimum, this sufficiency condition being stated in terms of explicit relationships that are to be satisfied by the observations and the variances and covariances of the variable and equation disturbances. This new result is especially significant for the applications since it can be used to verify whether or not a critical point is in fact a local minimum, a verification which is not to

be overlooked since this question cannot be resolved for the GLS-functional through a simple convexity argument as is done in the usual development of ordinary least squares theory when the errors are assumed to be independently distributed.

Finally we show that our results on estimating the parameters of a linear relationship having only "errors-in-variables" are general enough to include as special cases previously known and valid maximum likelihood or weighted least squares estimators for such stochastic structures derived by other authors under different assumptions. In particular we find that the celebrated Gauss-Markoff-Aitken [41,101,2] least squares estimators developed under the assumptions of uncorrelated or correlated errors in only the dependent variable can be obtained from our results. The works of Norton [113] and Parel [114] who worked with the linear regression model in which all variables are subject to errors is seen to be a special EV model, and their results can be written down immediately once this identification with our work on the EV model is made. Next the Kummel-Lindley [89,91] weighted least squares estimators for the parameters of a two variable linear relationship, derived under the specification that the errors-in-variables are uncorrelated and the ratio of their variances known, are seen to be a special case of these general results. Then the Koopmans-Tintner [84,144] maximum likelihood estimators for the parameters of a linear structure relationship in several variables in which all the variables are observed with errors which are intercorrelated within a multivariate observation, but not serially

correlated between observations, are obtained by adopting this specification in our analysis. As the last special case, we derive from our general results the recent work of C. Villegas [152], published in 1961 in the Annals of Mathematical Statistics in which he assumes that errors in all variables of a linear structure relationship are present, but not errors-in-equations, that these errors are not necessarily identically and independently normally distributed but rather that they may have intercorrelations and serial correlations of a nature whereby they jointly have a variance-covariance matrix which can be represented as the direct product of two known positive definite matrices.

1.4. PROBLEMS FOR FUTURE RESEARCH

We remark here that by no means have we given a unified exposition of the complete literature on the topic of the estimation problem for EV and EVE statistical models. Nor do we intend to imply by not having presented such a discussion of the many other important works on this topic in statistics that they are not necessarily to be preferred to the methods that have been mentioned. To do full justice to the literature and to explore the implications of other research topics connected with the estimation problem in EV and EVE statistical models would simply have carried us far beyond the scope of the present study.

Some of the problems which remain to be examined within the context of this work are the following: What properties such as weak or strong consistency or minimum variance do these ML- or GLS-estimators possess? Are there efficient computational algorithms for the numerical solution

of the non-linear normal equation systems that appear in this work? Can these results on estimating the unknown parameters of a single linear structural relationship having errors-in-variables and errors-in-equations be extended to multiple equation system in which such disturbances are jointly present? When prior information about the variances and covariances of the variable and equation errors is not available, can these additional unknown parameters be estimated jointly together with the parameters of the linear structural relationship if the model is identifiable? What are necessary and sufficient conditions for an EVE model to be identifiable? Or, when prior information about the parameters of an EV or EVE statistical model is available in the form of subjective or personalistic probability distributions [127], what are the implications for statistical inference of a Bayesian analysis?

Although we have not explored every aspect of the theory of estimation for EV and EVE statistical models, the contributions of our research to this topic cannot be placed in full perspective merely by a comparison of the several new results summarized in Section 1.3 vis-a-vis the problems which remain for future research. This research should also be viewed against the background of existing statistical literature concerning the theory of estimation in interdependent stochastic equation systems generally.

By reviewing the literature related to multiple equation systems where only serially independent equation errors are present and where non-linear equation systems for the estimators are also obtained [e.g.,

6,54,61,71,78,80,87,88,139,144,160], one finds that after about two decades of active research the problems listed above have not yet been fully resolved even for such interdependent stochastic equation systems. From this experience it appears that any one of these remaining problems for the EVE model involves a major investigation.

Thus the full contribution of our findings in this area of statistics stems not only from the several new results on this estimation problem which have been obtained, but from the following observations as well: First, the application of the Dwyer-Macphail matrix derivative calculus in this work clearly demonstrates the potential use of these same techniques for deriving the non-linear normal systems for the ML- or GLS-estimators of the parameters of a multiple equation EVE model whose variable and equation errors are both intercorrelated and serially correlated. Second, either by direct analysis of our non-linear system of normal equations for the general single equation EVE statistical model or by Monte Carlo techniques, the ML- or GLS-estimators can be expressed as functions of the variances and covariances of the disturbances, thus permitting the effects of variable or equation errors on these estimators to be evaluated more fully in a particular application than was previously possible. For these reasons, the author feels that the results of this study on estimation in EV and EVE statistical models provide new and useful contributions to statistics.

CHAPTER 2

EXTENSIONS OF THE DWYER-MACPHAIL MATRIX DERIVATIVE

2.1. INTRODUCTION

In this chapter we introduce higher-order Dwyer-Macphail matrix derivatives and then proceed to examine whether a scalar or matrix function of a matrix variable can be developed in a series expansion featuring its Dwyer-Macphail matrix derivatives. Before attempting a more complete presentation of these results, in order to provide the necessary insight and motivation, we first give a brief sketch of the elements of the calculus of Dwyer-Macphail matrix derivatives that had been developed before this research was undertaken.

2.2. THE DWYER-MACPHAIL MATRIX DERIVATIVE

Let Φ denote the real number system R or the complex number system C and

$$\Phi_{(m,n)} = \{X = [x_{ij}] | x_{ij} \in \Phi\}$$

the matrices of row order m and column order n with elements x_{ij} in Φ .

Consider a scalar function of a matrix variable, say

$$f: \Phi_{(m,n)} \rightarrow \Phi \quad \text{or} \quad y = f(X) .$$

The matrix space $\Phi_{(m,n)}$ together with scalar multiplication and matrix addition is an mn -dimensional vector space. Thus choosing a basis

In this connection we note the following

Theorem 2.2.1. If $y = f(X)$ is a well-defined function, in other words if $f(X_1) = f(X_2)$ whenever $X_1 = X_2$, then its representation $f(x_*)|B$ with respect to the basis B is a well-defined function too.

Proof. Suppose $y_1^* = f(x_*^{(1)}|B)$ and $y_2^* = f(x_*^{(2)}|B)$ where $x_*^{(1)} = x_*^{(2)}$. Since $x_*^{(1)} = x_*^{(2)}$, for every i and j , $x_{ij}^{(1)} = x_{ij}^{(2)}$ where $x_{ij}^{(k)}$ is the ij -th component of $x_*^{(k)}$ for $k = 1$ and 2 . Hence

$$X_1 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^{(1)} B_{ij} = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^{(2)} B_{ij} = X_2$$

and therefore

$$y_1^* = f(x_*^{(1)}|B) = f(X_1) = f(X_2) = f(x_*^{(2)}|B) = y_2^*.$$

We are now in a position to formally introduce the notion of the Dwyer-Macphail matrix derivative of a scalar function of a matrix variable. Consider a scalar function $y = f(X)$ of a matrix variable $X \in \Phi_{(m,n)}$ and its representation $f(x_*)|B$ with respect to the basis

$B = \{B_{ij} = [b_{rs}^{ij}]_{m \times n}\}$ for $\Phi_{(m,n)}$. Whenever the first partial derivatives $\partial f(x_*)|B / \partial x_{ij}^*$ exist for each $1 \leq i \leq m$ and $1 \leq j \leq n$ we say as in

Definition 2.2.2. The Dwyer-Macphail matrix derivative of $y = f(X)$ with respect to the basis B is the $m \times n$ -matrix,

$$\left[\frac{\partial f}{\partial X} \right]_B = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f(x_*|B)}{\partial x_{ij}^*} B_{ij} = \left[\sum_{i=1}^m \sum_{j=1}^n \frac{\partial f(x_*|B)}{\partial x_{ij}^*} b_{rs}^{ij} \right]_{m \times n}$$

$$B = \{B_{ij} = [b_{rs}^{ij}]_{m \times n} | i = 1, \dots, m \& j = 1, \dots, n\}$$

for $\Phi_{(m,n)}$ we may write each element $X \in \Phi_{(m,n)}$ uniquely as a linear combination of the $B_{ij} \in B$:

$$X = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^* B_{ij} = \left[\sum_{i=1}^m \sum_{j=1}^n x_{ij}^* b_{rs}^{ij} \right]_{m \times n} = [X]_B$$

where each x_{ij}^* is an element of Φ . Now the vector

$$x_* = (x_{11}^*, x_{12}^*, \dots, x_{1n}^*; \dots; x_{m1}^*, x_{m2}^*, \dots, x_{mn}^*)$$

whose components are the (unique) coefficients of X written as a linear combination of the elements B_{ij} of the basis B is a point in Φ^{mn} (the mn -dimensional cartesian product of Φ) and thus the function $y = f(X)$ from $\Phi_{(m,n)}$ to Φ (a scalar function of a matrix variable) may be represented by the function $f(x_*|B)$ from Φ^{mn} to Φ (a scalar function of several scalar variables) defined by the correspondence,

$$f(x_*|B) = f \left(\sum_{i=1}^m \sum_{j=1}^n x_{ij}^* B_{ij} \right) = f(X) .$$

Hence for a function $y = f(X)$ from $\Phi_{(m,n)}$ to Φ and a basis $B = \{B_{ij}\}$ for $\Phi_{(m,n)}$ we state the

Definition 2.2.1. The representation of $y = f(X)$ with respect to the basis $B = \{B_{ij}\}$ is that function from Φ^{mn} to Φ defined by the correspondence $f(x_*|B)$.

In this connection we note the following

Theorem 2.2.1. If $y = f(X)$ is a well-defined function, in other words if $f(X_1) = f(X_2)$ whenever $X_1 = X_2$, then its representation $f(x_*|B)$ with respect to the basis B is a well-defined function too.

Proof. Suppose $y_1^* = f(x_*^{(1)}|B)$ and $y_2^* = f(x_*^{(2)}|B)$ where $x_*^{(1)} = x_*^{(2)}$. Since $x_*^{(1)} = x_*^{(2)}$, for every i and j , $x_{ij}^{(1)} = x_{ij}^{(2)}$ where $x_{ij}^{(k)}$ is the ij -th component of $x_*^{(k)}$ for $k = 1$ and 2 . Hence

$$X_1 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^{(1)} B_{ij} = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^{(2)} B_{ij} = X_2$$

and therefore

$$y_1^* = f(x_*^{(1)}|B) = f(X_1) = f(X_2) = f(x_*^{(2)}|B) = y_2^*.$$

We are now in a position to formally introduce the notion of the Dwyer-Macphail matrix derivative of a scalar function of a matrix variable. Consider a scalar function $y = f(X)$ of a matrix variable $X \in \Phi_{(m,n)}$ and its representation $f(x_*|B)$ with respect to the basis

$B = \{B_{ij} = [b_{rs}^{ij}]_{m \times n}\}$ for $\Phi_{(m,n)}$. Whenever the first partial derivatives $\partial f(x_*|B) / \partial x_{ij}^*$ exist for each $1 \leq i \leq m$ and $1 \leq j \leq n$ we say as in

Definition 2.2.2. The Dwyer-Macphail matrix derivative of $y = f(X)$ with respect to the basis B is the $m \times n$ -matrix,

$$\left[\frac{\partial f}{\partial X} \right]_B = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f(x_*|B)}{\partial x_{ij}^*} B_{ij} = \left[\sum_{i=1}^m \sum_{j=1}^n \frac{\partial f(x_*|B)}{\partial x_{ij}^*} b_{rs}^{ij} \right]_{m \times n}$$

We refer to this matrix simply as the Dwyer-Macphail B matrix derivative of $y = f(X)$.

Usually in the applications that basis E whose elements are the $m \times n$ matrices E_{ij} with 1 in the ij -th place and 0 elsewhere is most convenient to use and the development of the Dwyer-Macphail matrix derivative was originally presented in [32] and later studied in [23] and [114] with this particular basis tacitly being assumed. In this case, suppressing any explicit notational reference to E we merely write

$$\left[\frac{\partial f}{\partial X} \right] = \left[\frac{\partial f(x)}{\partial x_{ij}} \right]_{m \times n} = \frac{\partial f}{\partial X}$$

and call $\partial f / \partial X$ the Dwyer-Macphail matrix derivative of $y = f(X)$ with respect to the basis E —or even more briefly—the Dwyer-Macphail matrix derivative of f . As we presently see there is no loss of generality or other disadvantage arising by restricting our attention to the Dwyer-Macphail matrix derivative $\partial f / \partial X$.

Suppose $B = \{B_{ij}\}$ and $\bar{B} = \{\bar{B}_{ij}\}$ are any two bases for the vector space $\Phi_{(m,n)}$. An element X of $\Phi_{(m,n)}$ has the representation

$$X = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^* B_{ij} = [X]_B$$

with respect to the basis B and the representation

$$X = \sum_{i=1}^m \sum_{j=1}^n \bar{x}_{ij}^* \bar{B}_{ij} = [X]_{\bar{B}}$$

with respect to the basis $\bar{B}$. Now the correspondence which associates the vector $x_* \in \Phi^{mxn}$ (whose components are the coefficients $\{x_{ij}^*\}$ of the representation of X with respect to the basis B) with the vector $\bar{x}_* \in \Phi^{mxn}$ (whose components are the coefficients $\{\bar{x}_{ij}^*\}$ of the representation of X with respect to the basis $\bar{B}$) is a well-defined function $\bar{x}_* = h(x_*)$ from Φ^{mxn} to Φ^{mxn} —for if $\bar{x}_*^{(1)} = h(x_*^{(1)})$ and $\bar{x}_*^{(2)} = h(x_*^{(2)})$ with $x_*^{(1)} = x_*^{(2)}$, then since the components $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ of $x_*^{(1)}$ and $x_*^{(2)}$ are equal,

$$X_1 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^{(1)} B_{ij} = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^{(2)} B_{ij} = X_2 ;$$

and, therefore, since an element of $\Phi_{(m,n)}$ has a unique representation with respect to any basis we must have $\bar{x}_*^{(1)} = \bar{x}_*^{(2)}$ or $h(x_*^{(1)}) = h(x_*^{(2)})$.

Next, turning our attention to the representations of a scalar function $y = f(X)$ of a matrix variable $X \in \Phi_{(m,n)}$ with respect to the bases B and $\bar{B}$, namely $f(x_*|B)$ and $f(\bar{x}_*|\bar{B})$ respectively, we find (provided the necessary partial derivatives exist), that

$$\frac{\partial f(x_*|B)}{\partial x_{ij}^*} = \frac{\partial f(\bar{x}_*|\bar{B})}{\partial \bar{x}_{ij}^*} \cdot \frac{\partial \bar{x}_{ij}^*}{\partial x_{ij}^*}$$

since $f(x_*|B) = f(\bar{x}_*|\bar{B})$ and $\bar{x}_* = h(x_*)$; or, in terms of the ij -th components, $\bar{x}_{ij}^* = h_{ij}(x_*)$. Thus we obtain the relation

$$\left[\frac{\partial f}{\partial X} \right]_B = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f(x_*|B)}{\partial x_{ij}^*} B_{ij} = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f(\bar{x}_*|\bar{B})}{\partial \bar{x}_{ij}^*} B_{ij}^* .$$

where

$$B_{ij}^* = \frac{\partial \bar{x}_{ij}^*}{\partial x_{ij}^*} B_{ij} .$$

Recalling that the Trace operator of a square matrix $A \in \Phi(k, k)$ is the sum of its diagonal elements or

$$\text{Tr } [A] = \sum_{i=1}^y a_{ii}$$

we now state

Theorem 2.2.2. If $[\partial f / \partial x]_B$ is the Dwyer-Macphail matrix derivative with respect to the basis $B = \{B_{ij}\}$ of the function $y = f(X)$ from $\Phi(m, n)$ to Φ and $\partial f / \partial X$ is its Dwyer-Macphail matrix derivative with respect to the basis $E = \{E_{ij}\}$, when both $[\partial f / \partial X]_B$ and $\partial f / \partial X$ exist, then

$$\left[\frac{\partial f}{\partial X} \right]_B = \sum_{i=1}^m \sum_{j=1}^n \text{Tr} \left(\frac{\partial f}{\partial X} E_{ij}^T \right) B_{ij}^* ,$$

where $B_{ij}^* = \frac{\partial x_{ij}}{\partial y_{ij}^*} B_{ij}$ and $x_{ij} = h_{ij}(y_*)$ or $x = h(y_*)$ is that function which represents the correspondence between the coefficients in the linear combinations expressing X in terms of the two bases B and E .

Proof. This theorem follows immediately from the expression for $f(x_* | B)$ when the basis B is taken to be the basis E and from the definition of the trace of a square matrix.

This result permits us to draw the important conclusion that there is no loss of generality in restricting our attention to $\partial f / \partial X$, the

Dwyer-Macphail matrix derivative with respect to the basis E , since $[\partial f / \partial X]_B$, the Dwyer-Macphail matrix derivative with respect to the basis B , can readily be computed from a knowledge of $\partial f / \partial X$ and the function $x = h(y_*)$ relating the vector variables x and y_* whose components are the coefficients of the linear combinations expressing an element X of $\Phi_{(m,n)}$ in terms of the basis E and the basis B respectively.

An important source of scalar functions of a matrix variable are the element functions of a matrix function of a matrix variable. In this connection consider a matrix function of a matrix variable from $\Phi_{(m,n)}$ to $\Phi_{(p,q)}$, say

$$F: \Phi_{(m,n)} \rightarrow \Phi_{(p,q)} \quad \text{or} \quad Y = F(X)$$

where $Y \in \Phi_{(p,q)}$ and $X \in \Phi_{(m,n)}$. Choosing a basis in $\Phi_{(m,n)}$,

$$B = \{B_{ij} = [b_{rs}^{ij}]_{m \times n} \mid i = 1, 2, \dots, m \& j = 1, 2, \dots, n\},$$

and a basis in $\Phi_{(p,q)}$,

$$C = \{C_{kl} = [c_{uv}^{kl}]_{p \times q} \mid k = 1, 2, \dots, p \& l = 1, 2, \dots, q\},$$

we may express $Y = F(X)$ in terms of the basis C and X in terms of the basis B :

$$[X]_B = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^* B_{ij}$$

and

$$[Y]_C = \sum_{k=1}^p \sum_{l=1}^q y_{kl}^* C_{kl} = \sum_{k=1}^p \sum_{l=1}^q f_{kl}(X|C) C_{kl} = [F(X)]_C.$$

In turn each element function $f_{kl}(X|C)$ in this representation of the matrix $Y = F(X)$ with respect to the basis C has (as a scalar function of a matrix variable) a representation $f_{kl}(x_*|B,C)$ with respect to the basis B , namely

$$y_{kl}^* = f_{kl}(x_*|B,C) = f_{kl}\left(\sum_{i=1}^m \sum_{j=1}^n x_{ij}^* B_{ij} \middle| C\right) = f_{kl}(X|C) .$$

In this manner we arrive at two representations for $Y = F(X)$,

$$\sum_{k=1}^p \sum_{l=1}^q f_{kl}(X|C) C_{kl} = F(X) = \sum_{k=1}^p \sum_{l=1}^q f_{kl}(x_*|B,C) C_{kl} ,$$

the right hand side being an expression for $F(X)$ which expresses it in terms of its element functions with respect to the basis C and the left hand side being an expression for $F(X)$ which expresses it in terms of the representations of these element functions in the basis C (considered as scalar functions of a matrix variable) with respect to the basis B .

Now we use the notation $\langle A \rangle_{uv}$ to denote the uv -th element of the matrix $A \in \Phi_{(p,q)}$; then if A is expressed as $[A]_C = [a_{kl}(C)]_{p \times q}$ in terms of the basis C for $\Phi_{(p,q)}$,

$$\langle A \rangle_{uv} = a_{uv}(C) .$$

Observing this notation, we find that for the matrix function $Y = F(X)$ expressed in terms of the basis C ,

$$\langle Y \rangle_{uv} = y_{uv}(C) = \sum_{k=1}^p \sum_{\ell=1}^q f_{k\ell}(X|C) c_{uv}^{k\ell} = \langle F(X) \rangle_{uv} ;$$

and as a scalar function of a matrix variable, it has a Dwyer-Macphail matrix derivative with respect to the basis B for $\Phi_{(m,n)}$, namely

$$\left[\frac{\partial \langle F(X) \rangle_{uv}}{\partial X} \right]_B = \left[\sum_{k=1}^p \sum_{\ell=1}^q \frac{\partial f_{k\ell}(x_*|B,C)}{\partial x_{rs}^*} c_{uv}^{k\ell} \right]_{mxn} .$$

Finally when we consider the basis D for $\Phi_{(p,q)}$ whose elements are the pxq matrices $D_{k\ell}$ (with 1 in the $k\ell$ -th place and 0 elsewhere) we merely write

$$Y = [y_{k\ell}] = [f_{k\ell}(X)] = F(X) ;$$

and for the Dwyer-Macphail matrix derivative of $\langle F(X) \rangle_{uv} = f_{uv}(X)$ with respect to the basis E for $\Phi_{(m,n)}$, we write

$$\frac{\partial \langle F(X) \rangle_{uv}}{\partial X} = \left[\frac{\partial f_{uv}(x)}{\partial x_{rs}} \right]_{mxn} = \frac{\partial f_{uv}(X)}{\partial X} .$$

Now that we have introduced these basic notions in the theory of Dwyer-Macphail matrix derivatives for scalar functions of a matrix variable and the notation that we are going to use for these mathematical ideas, in Section 2.3 we explore these notions in terms of particular matrix functions and later we indicate how the Dwyer-Macphail matrix derivative may be readily calculated for many functions by using several symbolic operational rules which relate the matrix derivative of a matrix function of a scalar variable to the Dwyer-Macphail matrix deriv-

ative of a scalar function of a matrix variable.

2.3. ILLUSTRATION OF DWYER-MACPHAIL MATRIX DIFFERENTIATION

We present an illustration of the ideas reviewed in Section 2.2 by deriving the Dwyer-Macphail matrix derivative of the element functions of a matrix function of a matrix variable arising from the extension of a scalar function of a scalar variable to the space of 2×2 matrices over the real or complex scalar field.

Accordingly let $f(\lambda)$ be any twice differentiable scalar function of a scalar variable from Φ to Φ ; e.g., $f(\lambda) = \lambda^n$, $f(\lambda) = e^\lambda$, $f(\lambda) = \sin \lambda$, etc. Now R. Rinehart [122] has shown that the Sylvester-Buchheim definition of a matrix function leads to the following extension of $f(\lambda)$ to a matrix function from and to the space of 2×2 matrices over the real or complex numbers: If $X \in \Phi_{(2,2)}$ and λ_1 and λ_2 are the eigenvalues of X , then the function $Y = F(X)$ which represents the extension of $y = f(x)$ to $\Phi_{(2,2)}$ is

$$F(X) = aX - bI \quad \text{when} \quad \lambda_1 \neq \lambda_2$$

or

$$F(X) = aX - bI \quad \text{when} \quad \lambda_1 = \lambda_2 = \lambda,$$

where I is the 2×2 identity matrix of $\Phi_{(2,2)}$ and

$$a = a(\lambda_1, \lambda_2) = \frac{f(\lambda_1) - f(\lambda_2)}{\lambda_1 - \lambda_2}$$

$$b = b(\lambda_1, \lambda_2) = \frac{\lambda_2 f(\lambda_1) - \lambda_1 f(\lambda_2)}{\lambda_1 - \lambda_2}$$

$$\begin{aligned} a^* = a^*(\lambda) &= \lim_{h \rightarrow 0} a(\lambda + h, \lambda - h) \\ &= f'(\lambda) \end{aligned}$$

$$\begin{aligned} b^* = b^*(\lambda) &= \lim_{h \rightarrow 0} b(\lambda + h, \lambda - h) \\ &= \lambda f'(\lambda) - f(\lambda) \end{aligned}$$

$$\begin{aligned} \lambda_1 = \lambda_1(x) &= \lambda_1(x_{11}, x_{12}, x_{21}, x_{22}) \\ &= \frac{x_{11} + x_{22}}{2} + \frac{1}{2} [(x_{11} - x_{22})^2 + 4x_{12}x_{21}]^{1/2} \end{aligned}$$

$$\begin{aligned} \lambda_2 = \lambda_2(x) &= \lambda_2(x_{11}, x_{12}, x_{21}, x_{22}) \\ &= \frac{x_{11} + x_{22}}{2} - \frac{1}{2} [(x_{11} - x_{22})^2 + 4x_{12}x_{21}]^{1/2} \end{aligned}$$

$$\begin{aligned} \lambda = \lambda(x) &= \lambda(x_{11}, x_{12}, x_{21}, x_{22}) \\ &= \frac{x_{11} + x_{22}}{2} \end{aligned}$$

From this explicit representation of $F(X)$ we can immediately calculate the Dwyer-Macphail matrix derivative of $y_{11} = \langle F(X) \rangle_{11}$, $y_{12} = \langle F(X) \rangle_{12}$, $y_{21} = \langle F(X) \rangle_{21}$, and $y_{22} = \langle F(X) \rangle_{22}$. Thus since

$$\langle F(X) \rangle_{ij} = ax_{ij} - b\delta_{ij} \quad \text{when} \quad \lambda_1 \neq \lambda_2$$

or

$$\langle F(X) \rangle_{ij} = a^*x_{ij} - b^*\delta_{ij} \quad \text{when} \quad \lambda_1 = \lambda_2 = \lambda$$

we have

$$\frac{\partial < F(X) >_{ij}}{\partial X} = \begin{bmatrix} \frac{\partial y_{ij}}{\partial x_{11}} & \frac{\partial y_{ij}}{\partial x_{12}} \\ \frac{\partial y_{ij}}{\partial x_{21}} & \frac{\partial y_{ij}}{\partial x_{22}} \end{bmatrix}$$

where, provided $\lambda_1 \neq \lambda_2$,

$$\frac{\partial y_{ij}}{\partial x_{rs}} = a \frac{\partial x_{ij}}{\partial x_{rs}} + x_{ij} \left(\sum_{k=1}^2 \frac{\partial a}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial x_{rs}} \right) - \delta_{ij} \left(\sum_{k=1}^2 \frac{\partial b}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial x_{rs}} \right);$$

or, provided $\lambda_1 = \lambda_2 = \lambda$,

$$\frac{\partial y_{ij}}{\partial x_{rs}} = a^* \frac{\partial x_{ij}}{\partial x_{rs}} + \frac{x_{ij}}{2} \frac{\partial (x_{11}+x_{22})}{\partial x_{rs}} \frac{da^*}{d\lambda} - \frac{\delta_{ij}}{2} \frac{\partial (x_{11}+x_{22})}{\partial x_{rs}} \frac{db^*}{d\lambda}.$$

As this example illustrates the calculation of the Dwyer-Macphail matrix derivative of a scalar function of a matrix variable by a direct application of Definition 2.2.2 can become extremely detailed. And it was this need to provide a more systematic way for finding the Dwyer-Macphail matrix derivative that certain rules were introduced into the calculus of Dwyer-Macphail matrix derivatives which had a symbolic interpretation and which enabled one to derive the Dwyer-Macphail matrix derivatives for a wide class of scalar functions of a matrix variable without requiring that a function of this class be explicitly expressed in terms of its representation as a scalar function of several scalar

variables, then having to compute all the partial derivatives of this function and arranging them into a matrix in accordance with the definition of the Dwyer-Macphail matrix derivative. It is our goal to develop a generalization of these rules by means of a duality type theorem between the Dwyer-Macphail matrix derivative of a scalar function of a matrix variable and the ordinary matrix derivative of a related matrix function of a scalar variable.

2.4. GENERALIZATION OF A USEFUL DUALITY TYPE THEOREM IN THE CALCULUS OF DWYER-MACPHAIL MATRIX DERIVATIVES

In this section we extend a basic result in the calculus of Dwyer-Macphail matrix derivatives which provided an ingenious method for deriving the Dwyer-Macphail matrix derivatives for a wide class of scalar functions of a matrix variable. This result was first established in [32] as a set of three rules which involved certain special matrices and their transposes. These rules could be used very abstractly or symbolically in the sense that one could interpret them as specifying when certain symbols were to be replaced by certain other symbols and how these symbols were to be arranged in order to obtain the Dwyer-Macphail matrix derivative. In this manner then these three rules made it possible to write down the Dwyer-Macphail matrix derivative for many scalar functions of a matrix variable quite routinely and mechanically once the ordinary matrix derivative of a related matrix function of a scalar variable was given.

Before developing the generalization of this key result in the calculus of Dwyer-Macphail matrix derivatives as a duality type theorem between the Dwyer-Macphail matrix derivative of a function of a matrix variable and the ordinary matrix derivative of a related matrix function of a scalar variable, however, we first explore some of the properties of the Trace operator.

The following notation is especially well suited for the purpose of presenting these properties of the Trace operator. For a matrix $A \in \Phi(m, n)$ we write

$$A = \begin{bmatrix} (A)_1 \\ (A)_2 \\ \vdots \\ (A)_m \end{bmatrix}$$

when we wish to consider the rows of A as row vectors

$$(A)_k = (a_{k1}, a_{k2}, \dots, a_{kn}) \in \Phi(1, n) ;$$

and we write

$$A = [(A)_1 \ (A)_2 \ \dots \ (A)_n]$$

when we wish to consider the columns of A as column vectors

$$\{A\}_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \cdot \\ \cdot \\ \cdot \\ a_{mk} \end{bmatrix} \in \Phi_{(m,1)}$$

We note that in this notation

$$A^T = \begin{bmatrix} (A^T)_1 \\ (A^T)_2 \\ \cdot \\ \cdot \\ (A^T)_n \end{bmatrix} \quad \text{and} \quad A^T = [\{A^T\}_1 \{A^T\}_2 \dots \{A^T\}_m]$$

where

$$(A^T)_k = \{A\}_k^T \quad \text{and} \quad \{A^T\}_k = (A)_k^T.$$

Keeping this notation in mind, consider $\lambda, \lambda_1, \lambda_2 \in \Phi$; $x, y \in \Phi_{(1,m)}$; $z \in \Phi_{(1,n)}$; $P, Q \in \Phi_{(m,m)}$; $U \in \Phi_{(m,n)}$; $V \in \Phi_{(n,m)}$; $A \in \Phi_{(r,s)}$; $B \in \Phi_{(s,u)}$; and $C \in \Phi_{(u,r)}$. An important theorem which we use over and over again throughout the remainder of this work is

Theorem 2.4.1. The Tr-operator enjoys the following properties:

$$2.4.1.1 \quad \lambda = \text{Tr } \lambda$$

$$2.4.1.2 \quad \text{Tr}(\lambda_1 P + \lambda_2 Q) = \lambda_1 \text{Tr} P + \lambda_2 \text{Tr} Q$$

$$2.4.1.3 \quad \text{Tr} P = \text{Tr} P^T$$

$$2.4.1.4 \quad \text{Tr} P P^T = \sum_{k=1}^m (P)_k (P)_k^T$$

$$2.4.1.5 \quad \text{Tr} UV = \text{Tr} VU = \text{Tr} U^T V^T = \text{Tr} V^T U^T$$

$$2.4.1.6 \quad \text{Tr}(x^T y) = \text{Tr}(y^T x)$$

$$2.4.1.7 \quad y U z^T = \text{Tr}(U^T y^T z)$$

$$2.4.1.8 \quad \text{Tr}(ABC) = \text{Tr}(B^T A^T C^T).$$

Proof. Properties 1 through 5 follow quite directly from the definition of the Tr-operator of a square matrix as the sum of the diagonal terms of the matrix. Property 6 follows from the fact that $xy^T = (xy^T)^T = yx^T$, hence from Property 1 that $\text{Tr} xy^T = \text{Tr} yx^T$. Finally applying Property 5 with $U = x e_{\Phi(1,m)}^T$ and $V = y^T e_{\Phi(m,1)}$ we have that

$$\text{Tr} y^T x = \text{Tr} xy^T = \text{Tr} yx^T = \text{Tr} x^T y.$$

The last two properties, Property 7 and Property 8, are not quite as trivial to prove. First we establish Property 8. Obviously

$$ABC = \begin{bmatrix} (A)_1 \\ (A)_2 \\ \vdots \\ (A)_r \end{bmatrix} B \begin{bmatrix} (C)_1 & (C)_2 & \dots & (C)_r \end{bmatrix}$$

or

$$ABC = \begin{bmatrix} (A)_1 B(C)_1 & (A)_1 B(C)_2 & \dots & (A)_1 B(C)_r \\ (A)_2 B(C)_1 & (A)_2 B(C)_2 & \dots & (A)_2 B(C)_r \\ \vdots & \vdots & \ddots & \vdots \\ (A)_r B(C)_1 & (A)_r B(C)_2 & \dots & (A)_r B(C)_r \end{bmatrix}.$$

Therefore, we find that

$$\begin{aligned} \text{Tr}(ABC) &= \sum_{k=1}^r (A)_k B(C)_k \\ &= \sum_{k=1}^r (A)_k B(C^T)_k^T \\ &= \sum_{k=1}^r \text{Tr}(B^T (A)_k^T (C^T)_k) \\ &= \text{Tr} \left(B^T \sum_{k=1}^r (A)_k^T (C^T)_k \right) \\ &= \text{Tr} \left(B^T \sum_{k=1}^r \{A^T\}_k (C^T)_k \right) \\ &= \text{Tr}(B^T A^T C^T). \end{aligned}$$

This last statement is merely a recognition of the fact that if $H \in \Phi(p, q)$

and $G \in \Phi(q, w)$ then

$$\begin{aligned}
 HG &= \left[\sum_{k=1}^q h_{ik} g_{kj} \right]_{pxw} \\
 &= \sum_{k=1}^q [h_{ik} g_{kj}]_{pxw} \\
 &= \sum_{k=1}^q (H)_k (G)_k .
 \end{aligned}$$

To establish Property 7 we first note that

$$\begin{aligned}
 U^T y^T z &= \begin{bmatrix} (U^T)_1 \\ (U^T)_2 \\ \vdots \\ (U^T)_n \end{bmatrix} [y^T z_1 \ y^T z_2 \ \dots \ y^T z_n] \\
 &= \begin{bmatrix} ((U^T)_1 y^T) z_1 & ((U^T)_1 y^T) z_2 & \dots & ((U^T)_1 y^T) z_n \\ ((U^T)_2 y^T) z_1 & ((U^T)_2 y^T) z_2 & \dots & ((U^T)_2 y^T) z_n \\ \vdots & \vdots & & \vdots \\ ((U^T)_n y^T) z_1 & ((U^T)_n y^T) z_2 & \dots & ((U^T)_n y^T) z_n \end{bmatrix}
 \end{aligned}$$

Therefore

$$\begin{aligned}
\text{Tr}(U^T y^T z) &= \sum_{k=1}^n ((U^T)_k y^T) z_k \\
&= \sum_{k=1}^n (\{U\}_k^T y^T) z_k \\
&= \sum_{k=1}^n (\{U\}_k^T y^T)^T z_k \\
&= \sum_{k=1}^n (y\{U\}_k) z_k \\
&= [y\{U\}_1 \ y\{U\}_2 \ \dots \ y\{U\}_n] z^T \\
&= y[\{U\}_1 \ \{U\}_2 \ \dots \ \{U\}_n] z^T \\
&= y \ U z^T .
\end{aligned}$$

With the proof of Theorem 2.4.1 completed, we are now in a position to derive a generalization of a significant result in the calculus of Dwyer-Macphail matrix derivatives which takes the form of a duality type theorem between Dwyer-Macphail matrix derivatives of scalar functions of a matrix variable and ordinary matrix derivatives of related matrix functions of a scalar variable. In what follows, whenever we refer to the bases E and D , respectively, for the matrix spaces $X \in \Phi_{(m,n)}$ and $Y \in \Phi_{(p,q)}$ in the context when these matrix spaces are the domain and range spaces of a matrix function of a matrix variable, $Y = F(X)$, in order to emphasize that the matrices of these bases are understood to have the same row and column orders as an independent matrix variable X and a matrix

function $Y = F(X)$, we use the notation J_{ij} to denote that member of the basis E for the domain space $\Phi_{(m,n)}$ of $F(X)$ with 1 in the ij -th place and 0 elsewhere and K_{rs} to denote that member of the basis D for the range space $\Phi_{(p,q)}$ of F with 1 in the rs -th place and 0 elsewhere.

Consider a matrix function of a scalar variable

$$H: \Phi \rightarrow \Phi_{(p,q)} \quad \text{or} \quad H(\lambda) = [h_{ij}(\lambda)]_{p \times q};$$

if each $h_{ij}(\lambda)$ is a differentiable function (in the usual sense in which differentiability of a scalar function of a scalar variable is understood) we say as in

Definition 2.4.1. The matrix function $H = H(\lambda) \in \Phi_{(p,q)}$ is a differentiable matrix function of the scalar variable $\lambda \in \Phi$ and the derivative of this matrix function of a scalar variable is defined as the matrix

$$\frac{dH}{d\lambda} = \left[\frac{dh_{ij}(\lambda)}{d\lambda} \right]_{p \times q}.$$

Moreover, $dH/d\lambda$ simply is called the ordinary (or scalar) matrix derivative of $H = H(\lambda)$.

Whenever we have a matrix function of a matrix variable, say

$$F: \Phi_{(m,n)} \rightarrow \Phi_{(p,q)} \quad \text{or} \quad Y = F(X),$$

by holding all elements of the matrix variable X constant except one, say x_{ij} (which we denote by λ), we obtain a matrix function of a scalar variable, namely

$$H(\lambda) = F(X_\lambda) = [f_{ij}(X_\lambda)]_{p \times q},$$

when X_λ denotes the matrix variable X when all of its elements are assumed to be held fixed or constant except one, x_{ij} , which is assumed to be a scalar variable $\lambda \in \Phi$. In this case we have

$$\frac{\partial H}{\partial \lambda} = \frac{\partial F(X_\lambda)}{\partial \lambda} = \left[\frac{\partial f_{ij}(x_\lambda)}{\partial \lambda} \right]_{p \times q}$$

where $f_{ij}(x_\lambda)$ is the representation of the ij -th element function $\langle F(X_\lambda) \rangle_{ij} = f_{ij}(X_\lambda)$ of $F(X_\lambda)$. In this case, however, we use the more direct notation

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = \frac{\partial F}{\partial \langle X \rangle_{ij}}$$

instead of

$$\frac{\partial H}{\partial \lambda} = \frac{\partial F(X_\lambda)}{\partial \lambda}.$$

Keeping these notions firmly in mind we can now state the following duality type theorem involving Dwyer-Macphail matrix derivatives and ordinary matrix derivatives, namely

Theorem 2.4.2. Consider a matrix function of a matrix variable,
 $Y = F(X)$, where $X \in \Phi_{(m,n)}$ and $Y \in \Phi_{(p,q)}$. Then

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \left[\text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{ij}} K_{rs}^T \right) \right]_{m \times n}$$

and

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = \left[\text{Tr} \left(\frac{\partial \langle Y \rangle_{rs}}{\partial X} J_{ij}^T \right) \right]_{pxq}$$

provided each of these matrix derivatives exists .

Proof. We first note that by definition the Dwyer-Macphail matrix derivative is given by

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \begin{bmatrix} \frac{\partial y_{rs}}{\partial x_{11}} & \frac{\partial y_{rs}}{\partial x_{12}} & \dots & \frac{\partial y_{rs}}{\partial x_{1n}} \\ \frac{\partial y_{rs}}{\partial x_{21}} & \frac{\partial y_{rs}}{\partial x_{22}} & \dots & \frac{\partial y_{rs}}{\partial x_{2n}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial y_{rs}}{\partial x_{m1}} & \frac{\partial y_{rs}}{\partial x_{m2}} & \dots & \frac{\partial y_{rs}}{\partial x_{mn}} \end{bmatrix},$$

while the ordinary matrix derivative is given by

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x_{ij}} & \frac{\partial y_{12}}{\partial x_{ij}} & \dots & \frac{\partial y_{1q}}{\partial x_{ij}} \\ \frac{\partial y_{21}}{\partial x_{ij}} & \frac{\partial y_{22}}{\partial x_{ij}} & \dots & \frac{\partial y_{2q}}{\partial x_{ij}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial y_{p1}}{\partial x_{ij}} & \frac{\partial y_{p2}}{\partial x_{ij}} & \dots & \frac{\partial y_{pq}}{\partial x_{ij}} \end{bmatrix}.$$

Hence we find by a direct analysis that

$$\text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{ij}} K_{rs}^T \right) = \frac{\partial y_{rs}}{\partial x_{ij}} = \text{Tr} \left(\frac{\partial \langle Y \rangle_{rs}}{\partial X} J_{ij}^T \right)$$

and the theorem follows immediately.

Let us consider for a moment the following matrix function of a matrix variable,

$$Y = AX ,$$

where $A \in \Phi_{(p,m)}$, $X \in \Phi_{(m,q)}$, and $Y \in \Phi_{(p,q)}$.

Then

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = AJ_{ij}$$

and, therefore,

$$\begin{aligned} \text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{ij}} K_{rs}^T \right) &= \text{Tr}((AJ_{ij})K_{rs}^T) \\ &= \text{Tr}(J_{ij}^T A^T K_{rs}) \\ &= \text{Tr}((A^T K_{rs})J_{ij}^T) \\ &= \langle A^T K_{rs} \rangle_{ij} . \end{aligned}$$

Thus from Theorem 2.4.2 we conclude that

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \left[\langle A^T K_{rs} \rangle_{ij} \right] = A^T K_{rs} .$$

Similarly, consider the matrix function of a matrix variable,

$$Y = XB ,$$

where $X \in \Phi_{(p,n)}$, $B \in \Phi_{(n,q)}$, and $Y \in \Phi_{(p,q)}$.

Then

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = J_{ij} B$$

and, therefore,

$$\begin{aligned} \text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{ij}} K_{rs}^T \right) &= \text{Tr}((J_{ij} B) K_{rs}^T) \\ &= \text{Tr}(B^T J_{ij}^T K_{rs}) \\ &= \text{Tr}((K_{rs} B^T) J_{ij}^T) \\ &= \langle K_{rs} B^T \rangle_{ij} . \end{aligned}$$

Again from Theorem 2.4.2 we conclude that

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \left[\langle K_{rs} B^T \rangle_{ij} \right] = K_{rs} B^T .$$

These examples illustrate an important theorem in the calculus of Dwyer-Macphail matrix derivatives, of which Theorem 2.4.2 is a generalization, which enables one to derive the Dwyer-Macphail matrix derivatives for many scalar functions of a matrix variable from the ordinary matrix derivatives of related matrix functions of a scalar variable by applying a few simple rules which can be given a symbolic interpretation. The proof of this theorem given below, as in the particular derivations for these two examples, is based directly on Theorem 2.4.1 and Theorem 2.4.2 and it is developed after the pattern of the derivations of these specific examples without need of new concepts or techniques. And it should be pointed out, moreover, that the proof of the following theorem

carried out along the lines suggested above is different from that given for the same theorem in [23] or [32].

We now state

Theorem 2.4.3. Suppose $\partial Y / \partial < X >_{ij}$ when expressed in terms of J_{ij} and J_{ij}^T has the form

$$\frac{\partial Y}{\partial < X >_{ij}} = \sum_{g=1}^G A_g J_{ij}^T B_g + \sum_{h=1}^H C_h J_{ij} D_h ,$$

in other words, that

$$Y = \sum_{g=1}^G A_g X^T B_g + \sum_{h=1}^H C_h X D_h = F(X)$$

where $A_g \in \Phi(p, n)$, $B_g \in \Phi(m, q)$, $C_h \in \Phi(p, m)$, and $D_h \in \Phi(n, q)$. Then $\partial < Y >_{rs} / \partial X$ can be written down immediately by

- a. changing each J_{ij} to K_{rs} and each J_{ij}^T to K_{rs}^T ,
- b. changing the pre (or post) multipliers of J_{ij} to their trans-
poses, and
- c. changing the pre (or post) multipliers of J_{ij}^T to post (or pre)
multipliers of K_{rs}^T .

Hence

$$\frac{\partial < Y >_{rs}}{\partial X} = \sum_{g=1}^G B_g K_{rs}^T A_g + \sum_{h=1}^H C_h^T K_{rs} D_h^T .$$

Conversely, $\partial Y / \partial < X >_{ij}$ can be obtained from $\partial < Y >_{rs} / \partial X$ by applying
the same rules interchanging J_{ij} and J_{ij}^T with K_{rs} and K_{rs}^T .

Proof. From the properties of the Tr-operator we find that

$$\begin{aligned} \text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{ij}} K_{rs}^T \right) &= \sum_{g=1}^G \text{Tr}(A_g J_{ij}^T B_g K_{rs}^T) \\ &+ \sum_{h=1}^H \text{Tr}(C_h J_{ij}^T D_h K_{rs}^T) . \end{aligned}$$

Now

$$\begin{aligned} \text{Tr}(A_g J_{ij}^T B_g K_{rs}^T) &= \text{Tr}((A_g J_{ij}^T)(B_g K_{rs}^T)) \\ &= \text{Tr}((B_g K_{rs}^T)(A_g J_{ij}^T)) \\ &= \text{Tr}((B_g K_{rs}^T A_g) J_{ij}^T) \\ &= \langle B_g K_{rs}^T A_g \rangle_{ij} \end{aligned}$$

$$\begin{aligned} \text{Tr}(C_h J_{ij}^T D_h K_{rs}^T) &= \text{Tr}(J_{ij}^T C_h^T)(K_{rs}^T D_h^T) \\ &= \text{Tr}(J_{ij}^T (C_h^T K_{rs}^T D_h^T)) \\ &= \text{Tr}((C_h^T K_{rs}^T D_h^T) J_{ij}^T) \\ &= \langle C_h^T K_{rs}^T D_h^T \rangle_{ij} . \end{aligned}$$

Thus we find that

$$\begin{aligned} \text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{ij}} K_{rs}^T \right) &= \sum_{g=1}^G \langle B_g K_{rs}^T A_g \rangle_{ij} \\ &+ \sum_{h=1}^H \langle C_h^T K_{rs}^T D_h^T \rangle_{ij} . \end{aligned}$$

Finally, applying Theorem 2.4.2, it follows that

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \sum_{g=1}^G \left[\langle B_g K_{rs}^T A_g \rangle_{ij} \right]_{m \times n} + \sum_{h=1}^H \left[\langle C_h^T K_{rs} D_h^T \rangle_{ij} \right]_{m \times n}$$

or, simply,

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \sum_{g=1}^G B_g K_{rs}^T A_g + \sum_{h=1}^H C_h^T K_{rs} D_h^T .$$

By reversing these steps we find that the converse of this theorem is also valid. Thus the proof of Theorem 2.4.3 is completed.

The following remarks make it evident that Theorem 2.4.3 (with a slight modification) can be used quite extensively as a basis for deriving Dwyer-Macphail matrix derivatives from ordinary matrix derivatives or vice versa. We note, first, that the conclusions of Theorem 2.4.3 are valid even when " $G = \infty$ " or " $H = \infty$ " provided that the elements of the corresponding (infinite) matrix series are finite. Now we know [60] that whenever $f(\lambda)$ is a (single-valued) function from Φ to Φ which is analytic in the sphere $|\lambda - \lambda_0| < r$, then the matrix function,

$$Y = F(X) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\lambda_0)}{k!} (X - \lambda_0 I)^k ,$$

whose domain and range is the Banach algebra of $m \times m$ matrices $\Phi_{(m,m)}$, is uniquely defined and analytic in the sphere $\|X - \lambda_0 I\| < r$ where $X \in \Phi_{(m,m)}$ and $\|A\|$ denotes the norm of the matrix $A \in \Phi_{(m,m)}$. For such matrix functions of a matrix variable we find that

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = \sum_{k=0}^{\infty} \frac{f^{(k)}(\lambda_0)}{k!} \frac{\partial (X - \lambda_0 I)^k}{\partial \langle X \rangle_{ij}},$$

provided the limiting operations can be interchanged.

Next, a direct computation shows that

$$\frac{\partial (X - \lambda_0 I)^k}{\partial \langle X \rangle_{ij}} = \sum_{s=0}^{k-1} (X - \lambda_0 I)^s J_{ij} (X - \lambda_0 I)^{k-s-1}.$$

For example, let

$$U = V^k \quad \text{and} \quad V = X - \lambda_0 I;$$

then $Z = (X - \lambda_0 I)^k$ is the composition of U with V . Let us choose an element $\langle X \rangle$ of the matrix variable X and denote this element by η ; similarly, denote the corresponding element $\langle V \rangle = \langle X \rangle - \lambda_0 \langle I \rangle$ of the matrix variable V by ξ . We also note that the matrix function Z_η of the scalar variable η (obtained by holding all elements of the matrix variable X constant except $\langle X \rangle$) is merely the composition of the matrix function U_ξ of the scalar variable ξ (obtained by holding all elements of the matrix variable V constant except $\langle V \rangle$) with the scalar function $\xi = g(\eta)$ (or $\langle V \rangle = \langle X \rangle - \lambda_0 \langle I \rangle$) of the scalar variable η ; in other words,

$$Z_\eta = U_\xi = V_\xi^k \quad \text{where} \quad \xi = g(\eta).$$

Thus from the "function of a function" differentiation theorem we have

that

$$\frac{\partial Z}{\partial \eta} = \frac{dU_\xi}{d\xi} \frac{d\xi}{d\eta}$$

or

$$\frac{\partial Z}{\partial \langle X \rangle} = \frac{\partial V^k}{\partial \langle V \rangle} \frac{d \langle V \rangle}{d \langle X \rangle},$$

since $\partial Z / \partial \eta$ is merely $\partial Z / \partial \langle X \rangle$ and $dU_{\xi} / d\xi$ is merely $\partial V^k / \partial \langle V \rangle$. Now from [32] we know that

$$\frac{\partial V^k}{\partial \langle V \rangle} = \sum_{s=0}^{k-1} V^s J V^{k-s-1};$$

and, of course,

$$\frac{d \langle V \rangle}{d \langle X \rangle} = 1.$$

Thus we obtain for $\partial Z / \partial \langle X \rangle$, where $Z = (X - \lambda_0 I)^k$,

$$\frac{\partial Z}{\partial \langle X \rangle} = \sum_{s=0}^{k-1} (X - \lambda_0 I)^s J (X - \lambda_0 I)^{k-s-1},$$

as had been indicated above.

Consequently for those matrix functions of a matrix variable defined as extensions of analytic scalar functions of a scalar variable in terms of infinite matrix series we find that

$$\frac{\partial Y}{\partial \langle X \rangle_{ij}} = \sum_{k=0}^{\infty} \frac{f^{(k)}(\lambda_0)}{k!} \sum_{s=0}^{k-1} (X - \lambda_0 I)^s J_{ij} (X - \lambda_0 I)^{k-s-1};$$

and by a direct application of Theorem 2.4.3 (with the modification noted) the Dwyer-Macphail matrix derivative of the function $Y = F(X)$ so defined is

$$\frac{\partial \langle Y \rangle_{rs}}{\partial X} = \sum_{k=0}^{\infty} \frac{f^{(k)}(\lambda_0)}{k!} \sum_{s=0}^{k-1} (X^T - \lambda_0 I)^s K_{rs} (X^T - \lambda_0 I)^{k-s-1}$$

where we recall that $(A^S)^T = (A^T)^S$.

Through this brief analysis, the scope of application of Theorem 2.4.3 and its usefulness in the calculus of Dwyer-Macphail matrix derivatives can be more fully appreciated.

For later applications of the Dwyer-Macphail matrix derivative involving estimation of unknown parameters in error-in-variable and error-in-equation statistical models, it is useful to note at this point that the Dwyer-Macphail matrix derivative of the matrix function $Z = (X - \lambda_0 I)^k$, rather than being obtained by a direct computation as presented in the above analysis, can be found by an application of a general theorem giving a Dwyer-Macphail matrix derivative "chain" rule for compound matrix functions of matrix variables, which for the calculus of Dwyer-Macphail matrix derivatives generalizes the usual "chain" rule of the ordinary calculus for differentiating compound scalar functions of scalar variables. Thus we state

Theorem 2.4.4. Consider a matrix function $Z = U(Y) \in \Phi_{(u,v)}$ of a matrix variable Y which in turn is related by another matrix function $Y = V(X) \in \Phi_{(p,q)}$ to the matrix variable $X \in \Phi_{(m,n)}$. Then the Dwyer-Macphail matrix derivative with respect to the index (i,j) of the matrix function $Z = U(V(X)) = W(X)$ is given by

$$\frac{\partial \langle W \rangle_{ij}}{\partial X} = \sum_{r=1}^p \sum_{s=1}^q \left\langle \frac{\partial \langle U \rangle_{ij}}{\partial Y} \right\rangle_{rs} \cdot \frac{\partial \langle V \rangle_{rs}}{\partial X}$$

for $1 \leq i \leq u$ and $1 \leq j \leq v$ provided the Dwyer-Macphail matrix derivatives with respect to the necessary indices of $Z = U(Y)$ and $Y = V(X)$ exist.

Proof. Writing

$$Z = U(Y) = [u_{ij}(y)]_{(i,j)}, \quad Y = V(X) = [v_{rs}(x)]_{(r,s)}$$

and

$$Z = W(X) = [w_{ij}(x)]_{(i,j)},$$

where, of course,

$$w_{ij}(x) = u_{ij}(y) \text{ with } y = (v_{11}(x), v_{12}(x), \dots, v_{pq}(x)),$$

upon forming the appropriate scalar partial derivatives appearing in the ordinary matrix derivative of $Z_\lambda = U(V(X_\lambda)) = W(X_\lambda)$ with respect to the element $\lambda = \langle X \rangle_{kl}$ of X in accordance with the usual chain rule for compound scalar functions of scalar variables, we obtain

$$\begin{aligned} \frac{\partial Z}{\partial \langle X \rangle_{kl}} &= \left[\frac{\partial w_{ij}(x_\lambda)}{\partial \lambda} \right]_{(i,j)} \\ &= \left[\sum_{r=1}^p \sum_{s=1}^q \frac{\partial u_{ij}(y)}{\partial y_{rs}} \cdot \frac{\partial v_{rs}(x)}{\partial \lambda} \right]_{(i,j)} \\ &= \sum_{r=1}^p \sum_{s=1}^q \frac{\partial v_{rs}(x)}{\partial \lambda} \left[\frac{\partial u_{ij}(y)}{\partial y_{rs}} \right]_{(i,j)} \\ &= \sum_{r=1}^p \sum_{s=1}^q \frac{\partial v_{rs}(x)}{\partial \lambda} \left(\frac{\partial U}{\partial \langle Y \rangle_{rs}} \right). \end{aligned}$$

Now, applying Theorem 2.4.2, we find that

$$\begin{aligned}
\frac{\partial \langle W \rangle_{ij}}{\partial X} &= \left[\text{Tr} \left(\frac{\partial Z}{\partial \langle X \rangle_{kl}} K_{ij}^T \right) \right]_{(k,l)} \\
&= \left[\sum_{r=1}^p \sum_{s=1}^q \frac{\partial v_{rs}(x)}{\partial \lambda} \text{Tr} \left(\frac{\partial U}{\partial \langle Y \rangle_{rs}} K_{ij}^T \right) \right]_{(k,l)} \\
&= \sum_{r=1}^p \sum_{s=1}^q \text{Tr} \left(\frac{\partial U}{\partial \langle Y \rangle_{rs}} K_{ij}^T \right) \left[\frac{\partial v_{rs}(x)}{\partial \lambda} \right]_{(k,l)} \\
&= \sum_{r=1}^p \sum_{s=1}^q \langle \frac{\partial \langle U \rangle_{ij}}{\partial Y} \rangle_{rs} \cdot \frac{\partial \langle V \rangle_{rs}}{\partial X}
\end{aligned}$$

and this completes the proof of Theorem 2.4.4.

As an immediate application of Theorem 2.4.4 we obtain the Dwyer-Macphail matrix derivative of $Z = (X - \lambda_0 I)^k$. Putting $Z = U(Y) = Y^k$ and $Y = V(X) = X - \lambda_0 I$ we see that $Z = W(X) = (X - \lambda_0 I)^k$ where $W(X) = U(V(X))$. Now.

$$\frac{\partial V}{\partial \langle X \rangle} = J \text{ and, consequently, } \frac{\partial \langle V \rangle}{\partial X} = K ;$$

thus, according to Theorem 2.4.4, we find that

$$\begin{aligned}
\frac{\partial \langle Z \rangle_{ij}}{\partial X} &= \sum_{r=1}^m \sum_{s=1}^m \text{Tr} \left(\frac{\partial v^k}{\partial \langle V \rangle_{rs}} K_{ij}^T \right) \cdot K_{rs} \\
&= \left[\text{Tr} \left(\frac{\partial v^k}{\partial \langle V \rangle_{rs}} K_{ij}^T \right) \right]_{(r,s)} \\
&= \frac{\partial \langle v^k \rangle_{ij}}{\partial V} ,
\end{aligned}$$

this last statement being a consequence of Theorem 2.4.2. And as we have indicated,

$$\frac{\partial \langle v^k \rangle_{ij}}{\partial v} = \sum_{s=0}^{k-1} (v^T)^s K_{ij} (v^T)^{k-s-1}$$

or

$$\frac{\partial \langle (X - \lambda_0 I)^k \rangle_{ij}}{\partial X} = \sum_{s=0}^{k-1} (X^T - \lambda_0 I)^s K_{ij} (X^T - \lambda_0 I)^{k-s-1},$$

the same result as obtained previously by a more direct computation.

2.5. HIGHER ORDER DWYER-MACPHAIL MATRIX DERIVATIVES

In this section we introduce several types of higher order matrix derivatives for matrix functions of a matrix variable; and in the course of this development, we obtain new results which correspond to the duality type theorem between Dwyer-Macphail matrix derivatives (of scalar functions of a matrix variable) and ordinary matrix derivatives (of matrix functions of a scalar variable) but which involve the higher order counterparts of these two types of first order matrix derivatives. These results permit us to derive such higher order matrix derivatives, one from the other, for many matrix functions, by repeated application of the same three simple rules given above in Theorem 2.4.3.

One of the five types of higher order matrix derivatives to be introduced below, namely the higher order Dwyer-Macphail matrix derivative, stands at the apex of a hierarchy of such constructions whose building blocks are the other four types of higher order matrix derivatives. In addition to being a natural generalization of the first order Dwyer-Macphail matrix derivative, it is in terms of these higher order Dwyer-

Macphail matrix derivatives that a series expansion for a matrix function of a matrix variable may be developed. We return to this point in Section 2.6; for the moment, however, we concentrate on presenting a theoretical development for these five new types of higher order matrix derivatives.

Thus let us consider a matrix function of a matrix variable, say

$$Y = F(X)$$

where $Y \in \Phi_{(p,q)}$ and $X \in \Phi_{(m,n)}$. In what follows we frequently write

$$X = \begin{bmatrix} x_{\gamma_1} & x_{\gamma_2} & \cdots & x_{\gamma_n} \\ x_{\gamma_{n+1}} & x_{\gamma_{n+2}} & \cdots & x_{\gamma_{2n}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{\gamma_{(m-1)n+1}} & x_{\gamma_{(m-1)n+2}} & \cdots & x_{\gamma_{mn}} \end{bmatrix}$$

and

$$Y = \begin{bmatrix} y_{\pi_1} & y_{\pi_2} & \cdots & y_{\pi_q} \\ y_{\pi_{q+1}} & y_{\pi_{q+2}} & \cdots & y_{\pi_{2q}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ y_{\pi_{(p-1)q+1}} & y_{\pi_{(p-1)q+2}} & \cdots & y_{\pi_{pq}} \end{bmatrix}$$

etc.; in other words, the ordered pair of integers (i,j) with $1 \leq i \leq m$ and $1 \leq j \leq n$ and (r,s) with $1 \leq r \leq p$ and $1 \leq s \leq q$ denoting the row and column positions of the elements of the matrix X and the matrix Y , respectively, have merely been relabeled in a way which emphasizes the cardinality of these matrix elements considered simply as a set of points, while giving secondary consideration to their row and column positions as elements of a matrix.

Keeping in mind the meaning of the indices γ_u for $u = 1, 2, \dots, mn$ and π_v for $v = 1, 2, \dots, pq$, while denoting the generic pair (i,j) with $1 \leq i \leq n$ and $1 \leq j \leq n$ by γ , and the generic pair (r,s) with $1 \leq r \leq p$ and $1 \leq s \leq q$ by π , and setting

$$\xi_1 = \pi, \xi_2 = \gamma_{m_1}, \dots, \xi_{k-1} = \gamma_{m_{k-2}}, \xi_k = \gamma_{m_{k-1}}$$

and

$$\eta_1 = \gamma_{n_1}, \eta_2 = \gamma_{n_2}, \dots, \eta_{k-1} = \gamma_{n_{k-1}}, \eta_k = \gamma,$$

we introduce the following types of higher order matrix derivatives for the matrix function $Y = F(X) \in \Phi_{(p,q)}$ of the matrix variable $X \in \Phi_{(m,n)}$:

The k -th order ordinary or scalar matrix derivative of $Y = F(X)$ with respect to the indices $\{\eta_1, \eta_2, \dots, \eta_k\}$ and denoted by

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}} ;$$

The k -th order partial Dwyer-Macphail matrix derivative of $Y = F(X)$ with respect to the indices $\{\xi_1, \xi_2, \dots, \xi_k\}$ and denoted by

$$\frac{\partial^k \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} ;$$

Two k-th order mixed partial matrix derivatives of $Y = F(X)$ with respect to the indices $\{\xi_1, \xi_2, \dots, \xi_{k-1}, \eta_{k-(i-1)}, \dots, \eta_k\}$ and $\{\xi_1, \xi_2, \dots, \xi_j, \eta_{j+1}, \dots, \eta_k\}$ and denoted by

$$\frac{\partial^i}{\partial X^{\eta_k} \dots \partial X^{\eta_{k-(i-1)}}} \left(\frac{\partial^{k-i} \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_{k-i})}{\partial X^{k-i}} \right)$$

and

$$\frac{\partial^j}{\partial X^j} \langle \frac{\partial^{k-j} Y}{\partial X^{\eta_k} \dots \partial X^{\eta_{j+1}}} \rangle (\xi_1, \xi_2, \dots, \xi_j) ;$$

Finally, the k-th order Dwyer-Macphail matrix derivative of $Y = F(X)$ with respect to the index π and denoted by

$$\frac{\partial^k \langle Y \rangle_\pi}{\partial X^k} .$$

Having introduced these five new types of higher order matrix derivatives by name together with the notation by which they are to be denoted, we now present a formal definition for each of these higher order matrix derivatives.

Definition 2.5.1. The k-th order ordinary or scalar matrix derivative of $Y = F(X)$ with respect to the indices $\{\eta_1, \eta_2, \dots, \eta_{k-1}, \eta_k\}$ is the pxq matrix

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_1} \partial \langle X \rangle_{\eta_2} \dots \partial \langle X \rangle_{\eta_1}} = \frac{\partial}{\partial \langle X \rangle_{\gamma}} \left(\frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}} \right)$$

where the notational convention

$$Y = \frac{\partial^0 Y}{\partial \langle X \rangle_0}$$

is adapted.

Theorem 2.5.1. If the k-th order partial derivatives of each of the element functions $y_{rs} = f_{rs}(x)$ of $Y = F(X)$ all exist then the k-th order ordinary matrix derivative of $Y = F(X)$ with respect to the indices $\{\eta_1, \eta_2, \dots, \eta_k\}$ exists and

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}} = \left[\frac{\partial^k f_{\pi}(x)}{\partial x_{\gamma} \partial x_{\gamma_{\eta_{k-1}}} \dots \partial x_{\gamma_{\eta_1}}} \right]_{\pi=(r,s)}.$$

Proof. This theorem is an immediate consequence of a direct examination of Definition 2.5.1 in which

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}}$$

is recursively defined in terms of

$$\frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}}.$$

Definition 2.5.2. The k-th order partial Dwyer-Macphail matrix derivative of $Y = F(X)$ with respect to the indices $\{\xi_1, \xi_2, \dots, \xi_{k-1}, \xi_k\}$ is the $m \times n$ matrix

$$\frac{\partial^k \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} = \frac{\partial}{\partial X} \left\langle \frac{\partial^{k-1} \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_{k-1})}{\partial X^{k-1}} \right\rangle_{\xi_k}$$

where the notational convention

$$Y = \frac{\partial^0}{\partial X^0} \langle Y \rangle$$

is understood.

Theorem 2.5.2. If the k-th order partial derivatives of the element function $y_{rs} = f_{rs}(x)$ of $Y = F(X)$ exist with respect to $x_{\xi_2}, x_{\xi_3}, \dots, x_{\xi_{k-1}}$ and x_{ij} in that order and for $1 \leq i \leq m$ and $1 \leq j \leq n$ then the k-th order partial Dwyer-Macphail matrix derivative with respect to the indices $\{\xi_1, \xi_2, \dots, \xi_k\}$ of $F(X)$ exists and

$$\frac{\partial^k \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} = \left[\frac{\partial^k f_{\pi}(x)}{\partial x_{\gamma} \partial x_{\gamma_{n_{k-1}}} \dots \partial x_{\gamma_{n_1}}} \right]_{\gamma=(i,j)}.$$

Proof. By direct reapplication of the recursive Definition 2.5.2 to the matrix derivative $\partial^{k-1} \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_{k-1}) / \partial X^{k-1}$ of order $k-1$ rather than k , we find that

$$\left\langle \frac{\partial^{k-1} \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_{k-1})}{\partial X^{k-1}} \right\rangle_{\xi_k} = \left\langle \frac{\partial}{\partial X} \left\langle \dots \frac{\partial}{\partial X} \left\langle \frac{\partial}{\partial X} \langle Y \rangle \right\rangle_{\xi_1} \right\rangle_{\xi_2} \dots \right\rangle_{\xi_{k-1}} \right\rangle_{\xi_k}$$

In turn we find that

$$\left\langle \frac{\partial}{\partial X} \left\langle \dots \frac{\partial}{\partial X} \left\langle \frac{\partial}{\partial X} \langle Y \rangle \right\rangle_{\xi_1} \right\rangle_{\xi_2} \dots \right\rangle_{\xi_{k-1}} \right\rangle_{\xi_k} = \frac{\partial^{k-1} f_{\pi}(x)}{\partial x_{\gamma_{n_{k-1}}} \partial x_{\gamma_{n_{k-2}}} \dots \partial x_{\gamma_{n_1}}};$$

consequently,

$$\frac{\partial^{k < Y > (\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} = \frac{\partial}{\partial X} \left(\frac{\partial^{k-1} f_{\pi}(x)}{\partial x_{\gamma_{n_{k-1}}} \partial x_{\gamma_{n_{k-2}}} \dots \partial x_{\gamma_{n_1}}} \right)$$

from which Theorem 2.5.2 can be established by a straightforward application of the notion of first order Dwyer-Macphail matrix derivative to the scalar function of a matrix variable

$$y = h(X) = \frac{\partial^{k-1} f_{\pi}(x)}{\partial x_{\gamma_{n_{k-1}}} \partial x_{\gamma_{n_{k-2}}} \dots \partial x_{\gamma_{n_1}}}$$

From these results a theorem directly analogous to Theorem 2.4.2 relating first order ordinary and Dwyer-Macphail matrix derivatives to each other can now be established for these higher order matrix derivatives, namely

Theorem 2.5.3. If $\eta_1 = \xi_2, \eta_2 = \xi_3, \dots, \eta_{k-1} = \xi_k$ and the necessary
k-th order partial derivatives exist then

$$\frac{\partial^{k < Y > (\pi, \xi_2, \dots, \xi_k)}{\partial X^k} = \left[\text{Tr} \left(\frac{\partial^k Y}{\partial < X >_{\gamma} \dots \partial < X >_{\eta_1}} K_{\pi}^T \right) \right]_{\gamma}$$

and

$$\frac{\partial^k Y}{\partial < X >_{\gamma} \partial < X >_{\eta_{k-1}} \dots \partial < X >_{\eta_1}} = \left[\text{Tr} \left(\frac{\partial^{k < Y > (\pi, \xi_2, \dots, \xi_k)}{\partial X^k} J_{\gamma}^T \right) \right]_{\pi}$$

Proof. Using Theorem 2.5.1 and Theorem 2.5.2, a direct computation shows that

$$\begin{aligned} \text{Tr} \left(\frac{\partial^k Y}{\partial x_{\eta_k} \partial x_{\eta_{k-1}} \dots \partial x_{\eta_1}} K_{\xi_1}^T \right) &= \frac{\partial^k y_{\pi}}{\partial x_{\eta_k} \partial x_{\eta_{k-1}} \dots \partial x_{\eta_1}} \\ &= \text{Tr} \left(\frac{\partial^k Y(\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} J_{\eta_k}^T \right); \end{aligned}$$

and from these considerations the theorem follows directly.

Unlike Theorem 2.4.2 relating first order ordinary and Dwyer-Macphail matrix derivatives, Theorem 2.5.3 by itself does not have the same direct import in the calculus of higher order matrix derivatives to the problem of deriving, generally, such higher order matrix derivatives from one another, since in the derivation of Theorem 2.5.3 we assumed that $\eta_1 = \xi_2$, $\eta_2 = \xi_3, \dots$, and $\eta_{k-1} = \xi_k$. In order to remove these restrictions and gain a more flexible result which allows us to derive each of these higher order matrix derivatives, one from the other, in a stepwise manner even when two given higher order matrix derivatives may be based on completely different indexing sets, we find it convenient to introduce the two types of mixed higher order matrix derivatives mentioned above.

Definition 2.5.3. The mixed k-th order partial ordinary matrix derivative of $Y = F(X)$ with respect to the indexing set $\{\xi_1, \dots, \xi_{k-i}, \eta_{k-(i-1)}, \dots, \eta_k\}$ is that matrix which is obtained by forming the i-th order ordinary matrix derivative with respect to the indices $\{\eta_{k-(i-1)}, \dots, \eta_k\}$ of the (k-i)-th order partial Dwyer-Macphail matrix derivative with respect to the indices $\{\xi_1, \dots, \xi_{k-i}\}$ of $Y = F(X)$; in

other words,

$$\frac{\partial^i}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_{k-(i-1)}}} \left(\frac{\partial^{k-i} \langle Y \rangle (\xi_1, \dots, \xi_{k-i})}{\partial X^{k-i}} \right) = \frac{\partial^i U}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_{k-(i-1)}}}$$

where

$$U = \frac{\partial^{k-i} \langle Y \rangle (\xi_1, \dots, \xi_{k-i})}{\partial X^{k-i}} = U(X) .$$

This mixed k-th order partial matrix derivative is defined for $i = 1, 2, \dots, k-1$ and accordingly our notation for this matrix derivative is meaningful only in these cases. By adopting the convention, however, that for $i = 0$ the undefined operator $\partial^i / \partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_{k-(i-1)}}$ is to be interpreted as $\partial^0 / \partial \langle X \rangle_0$ and that for $i = k$ the undefined operator $\frac{\partial^{k-i}}{\partial X^{k-i}} \langle Y \rangle (\xi_1, \dots, \xi_{k-i})$ is to be interpreted as $\partial^0 / \partial X^0$, when we recall the previously adopted convention that $Y = \frac{\partial^0}{\partial \langle X \rangle_0} Y$ and $Y = \frac{\partial^0}{\partial X^0} \langle Y \rangle$, our notation for this matrix derivative can be extended in the obvious manner for $i = 0$ or $i = k$ to include these cases.

Before discussing this mixed k-th order ordinary partial matrix derivative in detail, we first define the second type of mixed k-th order matrix derivative previously mentioned which we call the "mixed k-th order partial Dwyer-Macphail matrix derivative."

Definition 2.5.4. The mixed k-th order partial Dwyer-Macphail matrix derivative of $Y = F(X)$ with respect to the indexing set

$(\xi_1, \xi_2, \dots, \xi_j, \eta_{j+1}, \dots, \eta_k)$ is that matrix which is obtained by forming the j-th order partial Dwyer-Macphail matrix derivative with respect to

the indices $\{\xi_1, \xi_2, \dots, \xi_j\}$ of the $(k-j)$ -th order ordinary matrix deriva-
tive with respect to the indexing set $\{\eta_{j+1}, \dots, \eta_k\}$ of $Y = F(X)$; in
other words,

$$\frac{\partial^j}{\partial X^j} \left\langle \frac{\partial^{k-j} Y}{\partial X^{\eta_k} \dots \partial X^{\eta_{j+1}}} \right\rangle_{(\xi_1, \dots, \xi_j)} = \frac{\partial^j \langle V \rangle_{(\xi_1, \dots, \xi_j)}}{\partial X^j}$$

where

$$V = \frac{\partial^{k-j} Y}{\partial X^{\eta_k} \dots \partial X^{\eta_{j+1}}} = V(X) .$$

Again this mixed matrix derivative is defined by Definition 2.5.4 for $j = 1, 2, \dots, k-1$; and the notation introduced for it is meaningful only in these cases. But, exactly as in the case of the mixed matrix derivative defined by Definition 2.5.3, this notation may be extended to include the cases $j = 0$ or $j = k$. With this understood, we now move on to establish exactly how these two types of mixed matrix derivatives contribute to the development of a more general result than Theorem 2.5.3 which, as a by-product, allows us to derive either the k -th order partial Dwyer-Macphail matrix derivative or the k -th order ordinary matrix derivative, one from the other, even when these two higher order matrix derivatives are based on completely different indexing sets.

First, however, rather than state these general results immediately, let us explore their nature in a particular case. For the matrix function $Y = F(X) \in \Phi_{(p,q)}$ of the matrix variable $X \in \Phi_{(m,n)}$ we find by direct application of the various definitions involved that

$$\begin{aligned}
\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} &= \frac{\partial}{\partial X} \left\langle \frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \right\rangle_{\xi_2} \\
&= \frac{\partial}{\partial X} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\xi_2}} \right) \\
&= \left[\frac{\partial}{\partial \langle X \rangle_{\gamma}} \frac{\partial \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\xi_2}} \right]_{\gamma}
\end{aligned}$$

and

$$\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} = \left[\frac{\partial^2 \langle Y \rangle_{\pi}}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} \right]_{\pi} .$$

Now if $\eta_1 = \xi_2$ we find that

$$\frac{\partial^2 \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} = \text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right)$$

and, consequently,

$$\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} = \left[\text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_2 = \gamma} .$$

We also note that

$$\frac{\partial^2 \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} = \text{Tr} \left(\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} J_{\eta_2}^T \right)$$

and therefore,

$$\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} = \left[\text{Tr} \left(\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} J_{\eta_2}^T \right) \right]_{\xi_1 = \pi} .$$

These expressions for $\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2}$ in terms of $\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}$ and for $\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}$ in terms of $\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2}$, under the restric-

tion that $\eta_1 = \xi_2$, are in accord with the results of Theorem 2.5.3. But what if $\eta_1 \neq \xi_2$? Under these circumstances let us examine the 2-nd order mixed matrix derivatives $\frac{\partial}{\partial \langle X \rangle_{\eta_2}} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \right)$ and $\frac{\partial}{\partial X} \langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_2}} \rangle_{\xi_1}$.

Now

$$\begin{aligned} \frac{\partial}{\partial \langle X \rangle_{\eta_2}} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \right) &= \frac{\partial}{\partial \langle X \rangle_{\eta_2}} \left[\text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1} \\ &= \left[\text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1} . \end{aligned}$$

In turn,

$$\frac{\partial}{\partial \langle X \rangle_{\eta_2}} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\xi_2}} \right) = \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_2}} \frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} J_{\xi_2}^T \right) ;$$

and, therefore,

$$\begin{aligned} \frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2} &= \left[\frac{\partial^2 \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\xi_2}} \right]_{\eta_2} \\ &= \left[\text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_2}} \frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} J_{\xi_2}^T \right) \right]_{\eta_2} \\ &= \left[\text{Tr} \left[\text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1} J_{\xi_2}^T \right]_{\eta_2} . \end{aligned}$$

Thus when $\eta_1 \neq \xi_2$, via the mixed derivative $\frac{\partial}{\partial \langle X \rangle_{\eta_2}} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \right)$,

we have obtained a generalization of the previous relationship of Theorem 2.5.3 expressing $\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2}$ in terms of $\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}$

where we assumed that $\eta_1 = \xi_2$. Moreover, when in fact $\eta_1 = \xi_2$, we find that

$$\text{Tr} \left(\left[\text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1} J_{\eta_1}^T \right) = \text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right)$$

and the more general relationship reduces to the special relationship holding under this more restrictive assumption.

Next, we see in a direct fashion that,

$$\begin{aligned} \frac{\partial}{\partial X} \left\langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} &= \left[\frac{\partial}{\partial \langle X \rangle_{\xi_2}} \frac{\partial \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_2}} \right]_{\xi_2} \\ &= \left[\text{Tr} \left(\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} J_{\eta_2}^T \right) \right]_{\xi_2} \end{aligned}$$

provided the scalar 2-nd order partial derivatives appearing in these expressions are continuous. In turn,

$$\frac{\partial}{\partial \langle X \rangle_{\eta_1}} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_2}} \right) = \text{Tr} \left(\frac{\partial}{\partial X} \left\langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} J_{\eta_1}^T \right),$$

and, therefore,

$$\begin{aligned}
\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} &= \left[\frac{\partial^2 \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} \right]_{\xi_1} \\
&= \left[\text{Tr} \left(\frac{\partial}{\partial X} \langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_2}} \rangle_{\xi_1} J_{\eta_1}^T \right) \right]_{\xi_1} \\
&= \left[\text{Tr} \left[\text{Tr} \left(\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} J_{\eta_2}^T \right) \right]_{\xi_2} J_{\eta_1}^T \right]_{\xi_1}.
\end{aligned}$$

And again when $\eta_1 \neq \xi_2$, by using the mixed matrix derivative

$\frac{\partial}{\partial X} \langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_2}} \rangle_{\xi_1}$, we obtain a generalization of the previous relationship of Theorem 2.5.3 expressing $\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}$ in terms of $\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2}$ where we assumed that $\eta_1 = \xi_2$. Similarly when $\eta_1 = \xi_2$, we see that

$$\text{Tr} \left(\left[\text{Tr} \left(\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} J_{\eta_2}^T \right) \right]_{\xi_2} J_{\xi_2}^T \right) = \text{Tr} \left(\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} J_{\eta_2}^T \right)$$

and in this case too the more general result developed under the assumption that $\eta_1 \neq \xi_2$ reducing to the special result developed under the assumption that $\eta_1 = \xi_2$.

In summary, when $\eta_1 = \xi_2$, direct application of Theorem 2.5.3 permits us to derive higher order partial Dwyer-Macphail matrix derivatives from higher order ordinary matrix derivatives and vice versa through the following sequence of steps illustrated for second order matrix derivatives, namely:

$$\frac{\partial Y}{\partial \langle X \rangle_{\eta_1}} \rightarrow \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} \rightarrow \frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2}$$

and

$$\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \rightarrow \frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} \rightarrow \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}.$$

On the other hand, when $\eta_1 \neq \xi_2$, the above example suggests that by using mixed types of higher order matrix derivatives it is still possible to pass from one type of non-mixed higher order matrix derivative to the second type of non-mixed higher order matrix derivative through the re-application of the same kind of basic relationship among these derivatives as appears in Theorem 2.5.3 via a sequence of steps illustrated below for second order matrix derivatives:

$$\frac{\partial Y}{\partial \langle X \rangle_{\eta_1}} \rightarrow \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} \rightarrow \frac{\partial}{\partial \langle X \rangle_{\eta_2}} \left(\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \right) \rightarrow \frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2}$$

and

$$\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} \rightarrow \frac{\partial \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} \rightarrow \frac{\partial}{\partial X} \left\langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} \rightarrow \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}.$$

Keeping this discussion and the tableau above firmly in mind we now present a rigorous development of these notions which generalizes Theorem 2.4.2. We begin with

Theorem 2.5.4. Provided the necessary scalar k-th order partial derivatives exist and are continuous

$$\left\langle \frac{\partial^k \langle Y \rangle_{(\xi_1, \xi_2, \dots, \xi_k)}}{\partial X^k} \right\rangle_{\eta_k} = \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_k}} \frac{\partial^{k-1} \langle Y \rangle_{(\xi_1, \dots, \xi_{k-1})}}{\partial X^{k-1}} J_{\xi_k}^T \right);$$

and more generally for $i = 1, 2, \dots, k-1,$

$$\begin{aligned}
& \left\langle \frac{\partial^{k-(i-1)} \langle Y \rangle (\xi_1, \dots, \xi_{k-(i-1)})}{\partial X^{k-(i-1)}} \right\rangle_{\eta_{k-(i-1)}} \\
&= \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_{k-(i-1)}}} \frac{\partial^{k-i} \langle Y \rangle (\xi_1, \dots, \xi_{k-i})}{\partial X^{k-i}} J_{\xi_{k-(i-1)}}^T \right).
\end{aligned}$$

Proof. First we note as a consequence of Theorem 2.5.2 that

$$\frac{\partial^{k-1} \langle Y \rangle (\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} = \left[\frac{\partial^{k-1} \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta} \partial \langle X \rangle_{\xi_{k-1}} \dots \partial \langle X \rangle_{\xi_2}} \right]_{\eta}$$

Consequently,

$$\frac{\partial^{k-1} \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\xi_k} \partial \langle X \rangle_{\xi_{k-1}} \dots \partial \langle X \rangle_{\xi_2}} = \text{Tr} \left(\frac{\partial^{k-1} \langle Y \rangle (\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} J_{\xi_k}^T \right),$$

and the first part of this theorem follows immediately since

$$\begin{aligned}
\left\langle \frac{\partial^k \langle Y \rangle (\xi_1, \dots, \xi_k)}{\partial X^k} \right\rangle_{\eta_k} &= \frac{\partial^k \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\xi_k} \dots \partial \langle X \rangle_{\xi_2}} \\
&= \frac{\partial}{\partial \langle X \rangle_{\eta_k}} \left(\frac{\partial^{k-1} \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\xi_k} \dots \partial \langle X \rangle_{\xi_2}} \right) \\
&= \frac{\partial}{\partial \langle X \rangle_{\eta_k}} \text{Tr} \left(\frac{\partial^{k-1} \langle Y \rangle (\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} J_{\xi_k}^T \right) \\
&= \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_k}} \frac{\partial^{k-1} \langle Y \rangle (\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} J_{\xi_k}^T \right).
\end{aligned}$$

Now with an appropriate concern for the indices which appear in these relationships developed for the case $i = 1$, we see that the same arguments remain valid in the cases $i = 2, i = 3, \dots$, or $i = k-1$ too. Thus the proof of Theorem 2.5.4 can be completed in a direct fashion.

We now use the result of Theorem 2.5.4 to establish the following generalization of the first part of Theorem 2.5.3, namely

Theorem 2.5.5. Provided the required scalar k-th order partial derivatives exist and are continuous, then even without the restrictions that $\eta_1 = \xi_2$, $\eta_2 = \xi_3, \dots$, and $\eta_{k-1} = \xi_k$, the matrix derivative $\frac{\partial^k \langle Y \rangle(\xi_1, \dots, \xi_k)}{\partial X^k}$ may be expressed in terms of the matrix derivative $\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}}$ by the relation

$$\frac{\partial^k \langle Y \rangle(\xi_1, \dots, \xi_k)}{\partial X^k} = \left[\text{Tr} \left[\dots \left[\text{Tr} \left(\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1} \dots \right]_{\eta_{k-1}} J_{\xi_k}^T \right]_{\eta_k}.$$

Proof. Upon application of Theorem 2.5.4 we find that

$$\begin{aligned} \left\langle \frac{\partial^k \langle Y \rangle(\xi_1, \dots, \xi_k)}{\partial X^k} \right\rangle_{\eta_k} &= \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_k}} \frac{\partial^{k-1} \langle Y \rangle(\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} J_{\xi_k}^T \right) \\ \left\langle \frac{\partial^{k-1} \langle Y \rangle(\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} \right\rangle_{\eta_{k-1}} &= \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_{k-1}}} \frac{\partial^{k-2} \langle Y \rangle(\xi_1, \dots, \xi_{k-2})}{\partial X^{k-2}} J_{\xi_{k-1}}^T \right) \end{aligned}$$

etc., continuing to $i = k-1$, where we get

$$\left\langle \frac{\partial^2 \langle Y \rangle(\xi_1, \xi_2)}{\partial X^2} \right\rangle_{\eta_2} = \text{Tr} \left(\frac{\partial}{\partial \langle X \rangle_{\eta_2}} \frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} J_{\xi_2}^T \right).$$

Moreover, from Theorem 2.4.2 it follows that

$$\frac{\partial \langle Y \rangle_{\xi_1}}{\partial X} = \left[\text{Tr} \left(\frac{\partial Y}{\partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1}.$$

Finally, since

$$\frac{\partial^k \langle Y \rangle (\xi_1, \dots, \xi_k)}{\partial X^k} = \left[\frac{\partial^k \langle Y \rangle (\xi_1, \dots, \xi_k)}{\partial X^k} \right]_{\eta_k} \eta_k,$$

$$\frac{\partial^{k-1} \langle Y \rangle (\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} = \left[\frac{\partial^{k-1} \langle Y \rangle (\xi_1, \dots, \xi_{k-1})}{\partial X^{k-1}} \right]_{\eta_{k-1}} \eta_{k-1}$$

etc., for $i = 3, 4, \dots, k-1$, where for $i = k-1$,

$$\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2} = \left[\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2} \right]_{\eta_2} \eta_2,$$

upon successive substitution of the expressions for each of the elements of these matrices in terms of the Tr-operator of the product of the corresponding mixed partial ordinary matrix derivative and the transpose of an appropriate basis matrix, we obtain the desired expression and thus the proof of Theorem 2.5.5 is completed.

Thus Theorem 2.5.5 generalizes the first part of Theorem 2.5.3; and, moreover, if $\eta_1 = \xi_2, \dots, \eta_{k-1} = \xi_k$ we find that Theorem 2.5.5 reduces to the former Theorem 2.5.3 and becomes

$$\frac{\partial^k \langle Y \rangle (\xi_1, \dots, \xi_k)}{\partial X^k} = \left[\text{Tr} \left(\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_k}.$$

Next we undertake a generalization of the second part of Theorem 2.5.3; and, in so doing, we find the following theorem as useful in this task as Theorem 2.5.4 was in the generalization given of the first part of Theorem 2.5.3. Thus we state

Theorem 2.5.6. Provided the necessary scalar k-th order partial derivatives exist and are continuous,

$$\left\langle \frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} \right\rangle_{\xi_1} = \text{Tr} \left(\frac{\partial}{\partial X} \left\langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} J_{\eta_1}^T \right),$$

and more generally for $j = 1, 2, \dots, k-1$,

$$\begin{aligned} \left\langle \frac{\partial^{j-1}}{\partial X^{j-1}} \left\langle \frac{\partial^{k-(j-1)} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_j}} \right\rangle_{(\xi_1, \xi_2, \dots, \xi_{j-1})} \right\rangle_{\xi_j} \\ = \text{Tr} \left(\frac{\partial^j}{\partial X^j} \left\langle \frac{\partial^{k-j} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_{j+1}}} \right\rangle_{(\xi_1, \dots, \xi_j)} J_{\eta_j}^T \right). \end{aligned}$$

Proof. From Definition 2.5.4 we see that

$$\frac{\partial}{\partial X} \left\langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} = \left[\frac{\partial}{\partial \langle X \rangle_{\eta_1}} \frac{\partial^{k-1} \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right]_{\eta_1}.$$

Hence we readily obtain

$$\begin{aligned} \left\langle \frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} \right\rangle_{\xi_1} &= \frac{\partial^k \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} \\ &= \frac{\partial}{\partial \langle X \rangle_{\eta_1}} \left(\frac{\partial^{k-1} \langle Y \rangle_{\xi_1}}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right) \\ &= \text{Tr} \left(\frac{\partial}{\partial X} \left\langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} J_{\eta_1}^T \right). \end{aligned}$$

Similarly for $j = 2, 3, \dots, k-1$ we find applying Definition 2.5.4 that

$$\begin{aligned}
\frac{\partial^j}{\partial X^j} < \frac{\partial^{k-j} Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_{j+1}}} >_{(\xi_1, \dots, \xi_j)} \\
&= \frac{\partial}{\partial X} < \frac{\partial^{j-1}}{\partial X^{j-1}} < \frac{\partial^{k-j} Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_{j+1}}} >_{(\xi_1, \dots, \xi_{j-1})} >_{\xi_j} \\
&= \left[\frac{\partial}{\partial < X >_{\eta}} < \frac{\partial^{j-1}}{\partial X^{j-1}} < \frac{\partial^{k-j} Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_{j+1}}} >_{(\xi_1, \dots, \xi_{j+1})} >_{\xi_j} \right]_{\eta} ;
\end{aligned}$$

and as for the case $j = 1$ we again obtain

$$\begin{aligned}
&< \frac{\partial^{j-1}}{\partial X^{j-1}} < \frac{\partial^{k-(j-1)} Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_j}} >_{(\xi_1, \dots, \xi_{j-1})} >_{\xi_j} \\
&= \frac{\partial}{\partial < X >_{\eta_j}} < \frac{\partial^{j-1}}{\partial X^{j-1}} < \frac{\partial^{k-j} Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_{j+1}}} >_{(\xi_1, \dots, \xi_{j-1})} >_{\xi_j} \\
&= \text{Tr} \left(\frac{\partial^j}{\partial X^j} < \frac{\partial^{k-j} Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_{j+1}}} >_{(\xi_1, \dots, \xi_j)} J_{\eta_j}^T \right).
\end{aligned}$$

This then completes the proof of Theorem 2.5.6.

Using Theorem 2.5.6 we can readily establish the following important theorem:

Theorem 2.5.7. Provided the required scalar k-th order partial derivatives exist and are continuous, then even without the restrictions that $\eta_1 = \xi_2$, $\eta_2 = \xi_3, \dots$, and $\eta_{k-1} = \xi_k$, the matrix derivative

$$\frac{\partial^k Y}{\partial < X >_{\eta_k} \dots \partial < X >_{\eta_1}}$$

may be expressed in terms of the matrix derivative

$$\frac{\partial^k < Y >_{(\xi_1, \dots, \xi_k)}}{\partial X^k}$$

by the relation

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}}$$

$$= \left[\text{Tr} \left[\dots \left[\text{Tr} \left(\frac{\partial^k \langle Y \rangle(\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} J_{\eta_k}^T \right) \right] \dots \right] J_{\eta_1}^T \right] \xi_1.$$

Proof. In view of Theorem 2.5.6 we have

$$\left\langle \frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} \right\rangle_{\xi_1} = \text{Tr} \left(\frac{\partial}{\partial X} \left\langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} J_{\eta_1}^T \right)$$

$$\left\langle \frac{\partial}{\partial X} \left\langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \right\rangle_{\xi_1} \right\rangle_{\xi_2}$$

$$= \text{Tr} \left(\frac{\partial^2}{\partial X^2} \left\langle \frac{\partial^{k-2} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_3}} \right\rangle_{(\xi_1, \xi_2)} J_{\eta_2}^T \right)$$

etc., continuing to $j = k-1$ which gives

$$\left\langle \frac{\partial^{k-2}}{\partial X^{k-2}} \left\langle \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}}} \right\rangle_{(\xi_1, \dots, \xi_{k-2})} \right\rangle_{\xi_{k-1}}$$

$$= \text{Tr} \left(\frac{\partial^{k-1}}{\partial X^{k-1}} \left\langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_k}} \right\rangle_{(\xi_1, \dots, \xi_{k-1})} J_{\eta_{k-1}}^T \right).$$

In addition we see, putting $Z = \frac{\partial^{k-1}}{\partial X^{k-1}} \langle Y \rangle_{(\xi_1, \dots, \xi_{k-1})}$, that we may write the mixed matrix derivative which appears as a factor in the matrix product whose Trace is formed in the very last expression written above as follows:

$$\frac{\partial^{k-1}}{\partial X^{k-1}} \left\langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_k}} \right\rangle_{(\xi_1, \dots, \xi_{k-1})} = \frac{\partial}{\partial \langle X \rangle_{\eta_k}} Z;$$

and applying Theorem 2.4.2 to the right hand side of this expression we find that

$$\frac{\partial}{\partial \langle X \rangle_{\eta_k}} Z = \left[\text{Tr} \left(\frac{\partial \langle Z \rangle_{\xi_k}}{\partial X} J_{\eta_k}^T \right) \right]_{\xi_k}.$$

And since $\frac{\partial \langle Z \rangle_{\xi_k}}{\partial X} = \frac{\partial^k}{\partial X^k} \langle Y \rangle_{(\xi_1, \dots, \xi_k)}$, we obtain the relation

$$\frac{\partial^{k-1}}{\partial X^{k-1}} \langle \frac{\partial Y}{\partial \langle X \rangle_{\eta_k}} \rangle_{(\xi_1, \dots, \xi_{k-1})} = \left[\text{Tr} \left(\frac{\partial^k \langle Y \rangle_{(\xi_1, \dots, \xi_k)}}{\partial X^k} J_{\eta_k}^T \right) \right]_{\xi_k}.$$

Finally, noting that

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} = \left[\langle \frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} \rangle_{\xi_1} \right]_{\xi_1},$$

$$\frac{\partial}{\partial X} \langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \rangle_{\xi_1} = \left[\langle \frac{\partial}{\partial X} \langle \frac{\partial^{k-1} Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_2}} \rangle_{\xi_1} \rangle_{\xi_2} \right]_{\xi_2},$$

etc., continuing to $i = k-1$, where we have

$$\begin{aligned} \frac{\partial^{k-2}}{\partial X^{k-2}} \langle \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}}} \rangle_{(\xi_1, \dots, \xi_{k-2})} \\ = \left[\langle \frac{\partial^{k-2}}{\partial X^{k-2}} \langle \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}}} \rangle_{(\xi_1, \dots, \xi_{k-2})} \rangle_{\xi_{k-1}} \right]_{\xi_{k-1}}, \end{aligned}$$

upon successive substitution of the first set of expressions for each of the elements of these matrices, given in terms of the trace of the matrix product involving the corresponding mixed partial Dwyer-Macphail matrix derivatives and the transpose of an appropriate basis matrix, we obtain the desired expression thus completing the proof of this theorem.

We note for Theorem 2.5.7 too that it is a generalization of the

second part of Theorem 2.5.3 where we assumed that $\eta_1 = \xi_2, \dots$, and $\eta_{k-1} = \xi_k$; and Theorem 2.5.7 reduces to Theorem 2.5.3 under these restrictions giving us

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \dots \partial \langle X \rangle_{\eta_1}} = \left[\text{Tr} \left(\frac{\partial^k \langle Y \rangle (\xi_1, \dots, \xi_k)}{\partial X^k} J_{\eta_k}^T \right) \right]_{\xi_1}.$$

At this point it is instructive to review the calculus of higher order matrix derivatives (for those matrix derivatives defined in Definitions 2.5.1, 2.5.2, 2.5.3, and 2.5.4 as it has been developed in Theorems 2.5.3, 2.5.5, and 2.5.7) by applying these results to a matrix function of a matrix variable which is given as a product of constant and variable matrices. By so doing we arrive at a restatement of these theorems for a wide class of matrix functions of a matrix variable consisting of those matrix functions which are sums of products of constant and variable matrices. This result is analogous to Theorem 2.4.3 and allows us to derive these higher order matrix derivatives for such matrix functions by simply reapplying successively the same three basic rules stated in Theorem 2.4.3 for first order matrix derivatives of such matrix functions. The only modification of these rules necessary in applying them to the derivation of higher order matrix derivatives is that for higher order matrix derivatives like $\frac{\partial}{\partial \langle X \rangle} \left(\frac{\partial \langle Y \rangle}{\partial X} \right)$ the matrices J or J^T must be used in these rules in place of K or K^T to reflect the fact that this derivative has the same row and column orders as the matrix $\partial \langle Y \rangle / \partial X$ which, of course, are the row and column orders of X rather than Y ; while

for the first order matrix derivative $\partial Y / \partial \langle X \rangle$, the row and column orders are those of Y and hence K and K^T are used rather than J or J^T .

Consider the matrix variable $X \in \Phi(2,2)$ and the matrix function $Y = F(X) = XAX^T BX$ where A and B are constant matrices with $A, B \in \Phi(2,2)$ too. We denote X by

$$X = \begin{bmatrix} x_{\gamma_1} & x_{\gamma_2} \\ x_{\gamma_3} & x_{\gamma_4} \end{bmatrix}$$

and Y by

$$Y = \begin{bmatrix} y_{\pi_1} & y_{\pi_2} \\ y_{\pi_3} & y_{\pi_4} \end{bmatrix}$$

We find by a direct computation that

$$\frac{\partial Y}{\partial \langle X \rangle_{\eta_1}} = J_{\eta_1} A X^T B X + X A J_{\eta_1}^T B X + X A X^T B J_{\eta_1} ;$$

and

$$\begin{aligned} \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} &= J_{\eta_1} A J_{\eta_2}^T B X + J_{\eta_1} A X^T B J_{\eta_2} + J_{\eta_2} A J_{\eta_1}^T B X \\ &\quad + X A J_{\eta_1}^T B J_{\eta_2} + J_{\eta_2} A X^T B J_{\eta_1} + X A J_{\eta_2}^T B J_{\eta_1} ; \end{aligned}$$

and

$$\frac{\partial^3 Y}{\partial \langle X \rangle_{\eta_3} \partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} = J_{\eta_1} A J_{\eta_2}^T B J_{\eta_3} + J_{\eta_1} A J_{\eta_3}^T B J_{\eta_2} + J_{\eta_2} A J_{\eta_1}^T B J_{\eta_3} \\ + J_{\eta_3} A J_{\eta_1}^T B J_{\eta_2} + J_{\eta_2} A J_{\eta_3}^T B J_{\eta_1} + J_{\eta_3} A J_{\eta_2}^T B J_{\eta_1} ;$$

and, finally, for $k \geq 4$, that

$$\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}} = 0 \cdot I_2 .$$

Next applying Theorem 2.5.3 under the assumption that $\eta_1 = \xi_2$, $\eta_2 = \xi_3$, and $\eta_3 = \xi_4$, etc., and using the properties of the Tr-operator established in Theorem 2.4.1 we find that

$$\frac{\partial^2 \langle Y \rangle_{(\xi_1, \xi_2)}}{\partial X^2} = \left[\text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_2} \\ = \left[\text{Tr} (J_{\eta_1} A J_{\eta_2}^T B X K_{\xi_1}^T) \right]_{\eta_2} + \left[\text{Tr} (J_{\eta_1} A X^T B J_{\eta_2} K_{\xi_1}^T) \right]_{\eta_2} \\ + \left[\text{Tr} (J_{\eta_2} A J_{\eta_1}^T B X K_{\xi_1}^T) \right]_{\eta_2} + \left[\text{Tr} (X A J_{\eta_1}^T B J_{\eta_2} K_{\xi_1}^T) \right]_{\eta_2} \\ + \left[\text{Tr} (J_{\eta_2} A X^T B J_{\eta_1} K_{\xi_1}^T) \right]_{\eta_2} + \left[\text{Tr} (X A J_{\eta_2}^T B J_{\eta_1} K_{\xi_1}^T) \right]_{\eta_2} \\ = \left[\text{Tr} (B X K_{\xi_1}^T J_{\eta_1} A J_{\eta_2}^T) \right]_{\eta_2} + \left[\text{Tr} (J_{\eta_2}^T B^T X A^T J_{\eta_1}^T K_{\xi_1}) \right]_{\eta_2} \\ + \left[\text{Tr} (J_{\eta_2}^T K_{\xi_1} X^T B^T J_{\eta_1} A^T) \right]_{\eta_2} + \left[\text{Tr} (J_{\eta_2}^T B^T J_{\eta_1} A^T X^T K_{\xi_1}) \right]_{\eta_2} \\ + \left[\text{Tr} (J_{\eta_2}^T K_{\xi_1} J_{\eta_1}^T B^T X A^T) \right]_{\eta_2} + \left[\text{Tr} (B J_{\eta_1} K_{\xi_1}^T X A J_{\eta_2}^T) \right]_{\eta_2} \\ = B X K_{\xi_1}^T J_{\xi_2} A + B^T X A^T J_{\xi_2}^T K_{\xi_1} + K_{\xi_1} X^T B^T J_{\xi_2} A^T \\ + B^T J_{\xi_2} A^T X^T K_{\xi_1} + K_{\xi_1} J_{\xi_2}^T B^T X A^T + B J_{\xi_2} K_{\xi_1}^T X A ;$$

$$\begin{aligned}
\frac{\partial^3 \langle Y \rangle (\xi_1, \xi_2, \xi_3)}{\partial X^3} &= \left[\text{Tr} \left(\frac{\partial^3 Y}{\partial \langle X \rangle_{\eta_3} \partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_3} \\
&= \left[\text{Tr} (J_{\eta_1} A J_{\eta_2}^T B J_{\eta_3} K_{\xi_1}^T) \right]_{\eta_3} + \left[\text{Tr} (J_{\eta_1} A J_{\eta_3}^T B J_{\eta_2} K_{\xi_1}^T) \right]_{\eta_3} \\
&+ \left[\text{Tr} (J_{\eta_2} A J_{\eta_1}^T B J_{\eta_3} K_{\xi_1}^T) \right]_{\eta_3} + \left[\text{Tr} (J_{\eta_3} A J_{\eta_1}^T B J_{\eta_2} K_{\xi_1}^T) \right]_{\eta_3} \\
&+ \left[\text{Tr} (J_{\eta_2} A J_{\eta_3}^T B J_{\eta_1} K_{\xi_1}^T) \right]_{\eta_3} + \left[\text{Tr} (J_{\eta_3} A J_{\eta_2}^T B J_{\eta_1} K_{\xi_1}^T) \right]_{\eta_3} \\
&= \left[\text{Tr} (J_{\eta_3}^T B^T J_{\eta_2} A^T J_{\eta_1} K_{\xi_1}) \right]_{\eta_3} + \left[\text{Tr} (B J_{\eta_2} K_{\xi_1}^T J_{\eta_1} A J_{\eta_3}^T) \right]_{\eta_3} \\
&+ \left[\text{Tr} (J_{\eta_3}^T B^T J_{\eta_1} A^T J_{\eta_2} K_{\xi_1}) \right]_{\eta_3} + \left[\text{Tr} (J_{\eta_3}^T K_{\xi_1} J_{\eta_2}^T B^T J_{\eta_1} A^T) \right]_{\eta_3} \\
&+ \left[\text{Tr} (B J_{\eta_1} K_{\xi_1}^T J_{\eta_2} A J_{\eta_3}^T) \right]_{\eta_3} + \left[\text{Tr} (J_{\eta_3}^T K_{\xi_1} J_{\eta_1}^T B^T J_{\eta_2} A^T) \right]_{\eta_3} \\
&= B_{\xi_3}^T J_{\xi_3}^T A^T J_{\xi_2}^T K_{\xi_1} + B J_{\xi_3} K_{\xi_1}^T J_{\xi_2} A + B_{\xi_2}^T J_{\xi_2}^T A^T J_{\xi_3}^T K_{\xi_1} \\
&+ K_{\xi_1} J_{\xi_1}^T B^T J_{\xi_2} A^T + B J_{\xi_2} K_{\xi_1}^T J_{\xi_3} A + K_{\xi_1} J_{\xi_1}^T B^T J_{\xi_3} A^T ;
\end{aligned}$$

and, also, for $k \geq 4$,

$$\frac{\partial^k \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k} = \left[\text{Tr} \left(\frac{\partial^k Y}{\partial \langle X \rangle_{\eta_k} \partial \langle X \rangle_{\eta_{k-1}} \dots \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_k} = 0 \cdot I_2 .$$

Next, without making the assumptions that $\eta_1 = \xi_2$ or $\eta_2 = \xi_2$ or $\eta_3 = \xi_4$, etc., we derive these same matrix derivatives again, using the properties of the Tr-operator established in Theorem 2.4.1, but applying Theorem 2.5.5 instead of Theorem 2.5.3. Thus, for example,

$$\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2} = \left[\text{Tr} \left[\text{Tr} \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right]_{\eta_1} J_{\xi_2}^T \right]_{\eta_2} ;$$

now,

$$\begin{aligned}
& \left[\text{Tr} \left(\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right) \right]_{\eta_1} \\
&= \left[\text{Tr} (J_{\eta_1} A J_{\eta_2}^T B X K_{\xi_1}^T) \right]_{\eta_1} + \left[\text{Tr} (J_{\eta_1} A X^T B J_{\eta_2} K_{\xi_1}^T) \right]_{\eta_1} + \left[\text{Tr} (J_{\eta_2} A J_{\eta_1}^T B X K_{\xi_1}^T) \right]_{\eta_1} \\
&+ \left[\text{Tr} (X A J_{\eta_1}^T B J_{\eta_2} K_{\xi_1}^T) \right]_{\eta_1} + \left[\text{Tr} (J_{\eta_2} A X^T B J_{\eta_1} K_{\xi_1}^T) \right]_{\eta_1} + \left[\text{Tr} (X A J_{\eta_2}^T B J_{\eta_1} K_{\xi_1}^T) \right]_{\eta_1} \\
&= \left[\text{Tr} (J_{\eta_1}^T K_{\xi_1} X^T B^T J_{\eta_2} A^T) \right]_{\eta_1} + \left[\text{Tr} (J_{\eta_1}^T K_{\xi_1} J_{\eta_2}^T B^T X A^T) \right]_{\eta_1} + \left[\text{Tr} (B X K_{\xi_1}^T J_{\eta_2} A J_{\eta_1}^T) \right]_{\eta_1} \\
&+ \left[\text{Tr} (B^T J_{\eta_2} K_{\xi_1}^T X A J_{\eta_1}^T) \right]_{\eta_1} + \left[\text{Tr} (J_{\eta_1}^T B^T X A^T J_{\eta_2}^T K_{\xi_1}) \right]_{\eta_1} + \left[\text{Tr} (J_{\eta_1}^T B^T J_{\eta_2} A^T X^T K_{\xi_1}) \right]_{\eta_1} \\
&= K_{\xi_1} X^T B^T J_{\eta_2} A^T + K_{\xi_1} J_{\eta_2}^T B^T X A^T + B X K_{\xi_1}^T J_{\eta_2} A + B J_{\eta_2} K_{\xi_1}^T X A \\
&+ B^T X A^T J_{\eta_2}^T K_{\xi_1} + B^T J_{\eta_2} A^T X^T K_{\xi_1}
\end{aligned}$$

In turn, we find that

$$\begin{aligned}
& \left[\text{Tr} \left[\text{Tr} \frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}} K_{\xi_1}^T \right]_{\eta_1} J_{\xi_2}^T \right]_{\eta_2} \\
&= \left[\text{Tr} (K_{\xi_1} X^T B^T J_{\eta_2} A^T J_{\xi_2}^T) \right]_{\eta_2} + \left[\text{Tr} (K_{\xi_1} J_{\eta_2}^T B^T X A^T J_{\xi_2}^T) \right]_{\eta_2} \\
&+ \left[\text{Tr} (B X K_{\xi_1}^T J_{\eta_2} A J_{\xi_2}^T) \right]_{\eta_2} + \left[\text{Tr} (B J_{\eta_2} K_{\xi_1}^T X A J_{\xi_2}^T) \right]_{\eta_2} \\
&+ \left[\text{Tr} (B^T X A^T J_{\eta_2}^T K_{\xi_1} J_{\xi_2}^T) \right]_{\eta_2} + \left[\text{Tr} (B^T J_{\eta_2} A^T X K_{\xi_1} J_{\xi_2}^T) \right]_{\eta_2} \\
&= \left[\text{Tr} (J_{\eta_2}^T B X K_{\xi_1}^T J_{\xi_2} A) \right]_{\eta_2} + \left[\text{Tr} (B^T X A^T J_{\xi_2}^T K_{\xi_1} J_{\eta_2}^T) \right]_{\eta_2} \\
&+ \left[\text{Tr} (J_{\eta_2}^T K_{\xi_1} X^T B^T J_{\xi_2} A^T) \right]_{\eta_2} + \left[\text{Tr} (J_{\eta_2}^T B^T J_{\xi_2} A^T X^T K_{\xi_1}) \right]_{\eta_2} \\
&+ \left[\text{Tr} (K_{\xi_1} J_{\xi_2}^T B^T X A^T J_{\eta_2}^T) \right]_{\eta_2} + \left[\text{Tr} (J_{\eta_2}^T B J_{\xi_2} K_{\xi_1}^T X A) \right]_{\eta_2} \\
&= B X K_{\xi_1}^T J_{\xi_2} A + B^T X A^T J_{\xi_2}^T K_{\xi_1} + K_{\xi_1} X^T B^T J_{\xi_2} A^T \\
&+ B^T J_{\xi_2} A^T X^T K_{\xi_1} + K_{\xi_1} J_{\xi_2}^T B^T X A^T + B J_{\xi_2} K_{\xi_1}^T X A .
\end{aligned}$$

Thus we see that under the assumption that $\eta_1 \neq \xi_2$, an application of Theorem 2.5.5 leads us to the same result as that obtained through an application of Theorem 2.5.3 under the assumption that $\eta_1 = \xi_2$. Moreover, under the further assumption that $\eta_2 \neq \xi_3$ and $\eta_3 \neq \xi_4$, a similar application of Theorem 2.5.5, but involving three and four rather than two reapplications of the same type of basic derivations as that just given for $\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2}$ when $\eta_1 \neq \xi_2$ would lead us to the same expressions for $\frac{\partial^3 \langle Y \rangle (\xi_1, \xi_2, \xi_3)}{\partial X^3}$ and $\frac{\partial^4 \langle Y \rangle (\xi_1, \xi_2, \xi_3, \xi_4)}{\partial X^4}$ as previously developed by application of Theorem 2.5.3 when it was assumed that $\eta_2 = \xi_3$ and $\eta_3 = \xi_4$. The verification of this statement is omitted, however.

In turn by applying the second part of Theorem 2.5.3 (under the assumption that $\eta_1 = \xi_2$, $\eta_2 = \xi_3$, and $\eta_3 = \xi_4$) or Theorem 2.5.7 (under the assumption that $\eta_1 \neq \xi_2$, $\eta_2 \neq \xi_3$, and $\eta_3 \neq \xi_4$) and using the same properties of the Tr-operator as were used above, we may derive the matrix derivatives

$$\frac{\partial^2 Y}{\partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}, \quad \frac{\partial^3 Y}{\partial \langle X \rangle_{\eta_3} \partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}$$

and

$$\frac{\partial^4 Y}{\partial \langle X \rangle_{\eta_4} \partial \langle X \rangle_{\eta_3} \partial \langle X \rangle_{\eta_2} \partial \langle X \rangle_{\eta_1}}$$

from the matrix derivatives

$$\frac{\partial^2 \langle Y \rangle (\xi_1, \xi_2)}{\partial X^2}, \quad \frac{\partial^3 \langle Y \rangle (\xi_1, \xi_2, \xi_3)}{\partial X^3} \quad \text{and} \quad \frac{\partial^4 \langle Y \rangle (\xi_1, \xi_2, \xi_3, \xi_4)}{\partial X^4} .$$

We omit these computations, however, since they proceed in the same manner as those already given above taking into consideration, of course, the modifications implied by the two Theorems 2.5.3 and 2.5.7. In fact, this example is merely an illustration of the following theorem which is a re-statement of these results for matrix functions of a matrix variable given by sums of products of constant and variable matrices, namely

Theorem 2.5.8. If the matrix function $Y = F(X)$ of the matrix variable X is a sum of products of constant matrices and variable matrices X or X^T , then the matrix derivatives $\frac{\partial^k \langle Y \rangle (\xi_1, \xi_2, \dots, \xi_k)}{\partial X^k}$ and $\frac{\partial^k Y}{\partial X^k}$ can be derived, one from the other, by repeated reapplication of the three simple rules previously established for the first order derivatives $\partial \langle Y \rangle_\xi / \partial X$ and $\partial Y / \partial \langle X \rangle_\eta$ of such matrix functions, the only modification required of these rules for higher order matrix derivatives being that, except for the first step, the matrices J and J^T appropriately indexed are to be used in place of the matrices K or K^T similarly indexed in each subsequent step.

This theorem follows directly as an immediate consequence of the application of Theorem 2.5.3 or Theorem 2.5.5 and Theorem 2.5.7 and the properties of the Tr-operator to any matrix function given as a sum of products of constant and variable matrices as illustrated in the example for the matrix functions $Y = F(X) = XAX^T BX$ with $X, A, B \in \Phi_{(2,2)}$ which we discussed above in detail.

2.6. SERIES EXPANSIONS FEATURING DWYER-MACPHAIL MATRIX DERIVATIVES

Before introducing our fifth and last type of higher order matrix derivative, the k -th order Dwyer-Macphail matrix derivative, a matrix derivative which generalizes the first order Dwyer-Macphail matrix derivative and in terms of which a matrix function of a matrix variable may be expressed as an (infinite) matrix series expansion featuring its first and higher order Dwyer-Macphail matrix derivatives, we need to introduce first a few preliminary notions which facilitate this investigation. Specifically, we need to recall the notion of a "direct product" of matrices which has been explored by other authors as, for example, by Halmos [55].

Definition 2.6.1. The (right) direct product $A \otimes B$ of $A = [a_{ij}] \in \Phi(m, n)$ and $B = [b_{rs}] \in \Phi(p, q)$ is that $mp \times nq$ matrix obtained by replacing each element a_{ij} of A by the matrix $a_{ij} \cdot B$ as a block; in other words,

$$A \otimes B = [a_{ij}]_{(i,j)} \otimes B = [a_{ij} \cdot B]_{(i,j)} = [a_{ij} [b_{rs}]_{(r,s)}]_{(i,j)} = [[a_{ij} b_{rs}]_{(r,s)}]_{(i,j)}$$

The following theorem is useful:

Theorem 2.6.1. If $A \in \Phi(m, n)$ and $B \in \Phi(p, q)$ then $(A \otimes B)^T = A^T \otimes B^T$; and more generally, for any positive integer k , if

$$B_1 \in \Phi(m_1, n_1), B_2 \in \Phi(m_2, n_2), \dots, B_k \in \Phi(m_k, n_k)$$

then

$$(B_1 \otimes B_2 \otimes \dots \otimes B_k)^T = B_1^T \otimes B_2^T \otimes \dots \otimes B_k^T.$$

Proof. Since $A \otimes B = [a_{ij} \cdot B]$ it follows that $(A \otimes B)^T = [a_{ij} \cdot B]^T = [a_{ji} \cdot B^T] = A^T \otimes B^T$. Now assume that the theorem holds for a particular positive integer k and let us consider $B_1 \otimes B_2 \otimes \dots \otimes B_{k+1}$. Putting $A_k = B_1 \otimes \dots \otimes B_k$ it follows that $(B_1 \otimes \dots \otimes B_{k+1})^T = (A_k \otimes B_{k+1})^T = A_k^T \otimes B_{k+1}^T = (B_1 \otimes \dots \otimes B_k)^T \otimes B_{k+1}^T = B_1^T \otimes B_2^T \otimes \dots \otimes B_k^T \otimes B_{k+1}^T$ by virtue of our induction hypothesis. We have demonstrated, therefore, that if the proposition holds for k then it holds for $k+1$ too; and knowing that the theorem is valid for the direct product of two matrices, we may conclude by mathematical induction that the theorem is valid for any (finite) number of factors.

Whenever $B_1 = B_2 = \dots = B_k = H \in \Phi_{(m,n)}$, rather than writing $B_1 \otimes B_2 \otimes \dots \otimes B_k$ we merely write H^k as a shorter notation for denoting the direct product of H with itself k -times. Although, when H is a square matrix, this interpretation of the notation H^k conflicts with an established meaning that H^k denotes the matrix product (rather than the direct product) of H with itself k -times, in the specialized context in which the symbol H^k appears in this work, there need be no confusion or ambiguity as to which meaning is intended for the notation H^k when H is a square matrix. Thus, when H is a square matrix, it will be quite clear from the context in which the notation H^k appears in this work that this notation is to be interpreted as the direct product—not the matrix product—of H with itself k -times.

Moreover, adapting the notation H^k to denote the direct product of H with itself k -times, when the k matrix factors $B_1, B_2, \dots$, and B_k ap-

peating in Theorem 2.6.1 are all identical, this theorem can simply be restated in the simple form

$$(H^k)^T = (H^T)^k$$

where H denotes the common matrix value of the k matrix factors

$B_1, B_2, \dots, B_k$ appearing in this theorem.

Having disposed of these few preliminary, but essential notions, we now find ourselves in a position to introduce

Definition 2.6.2. For a matrix $X \in \Phi_{(m,n)}$ let us take as the mn -matrix operator $\partial/\partial X$ that matrix obtained by replacing each element $\langle X \rangle_{ij}$ of X by the scalar partial differentiation operator $\partial/\partial \langle X \rangle_{ij}$. Then the k -th order Dwyer-Macphail matrix derivative with respect to the index $\pi = (r,s)$ of the matrix function $Y = F(X) \in \Phi_{(p,q)}$ of the matrix variable $X \in \Phi_{(m,n)}$ is the $mp \times nq$ -matrix

$$\frac{\partial^k \langle Y \rangle_\pi}{\partial X^k} = \frac{\partial}{\partial X} \otimes \frac{\partial}{\partial X} \otimes \dots \otimes \frac{\partial}{\partial X} \langle Y \rangle_\pi$$

where the matrix operator $\partial/\partial X$ is understood to appear k -times on the right hand side in this expression.

Recalling that for a matrix $X \in \Phi_{(m,n)}$ we have labeled by $\gamma_k = (i_k, j_k)$ for $k = 1, 2, \dots, mn$ the ordered pairs whose components are the row and column indices, respectively, of the element $\langle X \rangle_{i_k j_k}$ of X , this labeling being done from left to right in each successive row of X , let us consider the matrix whose elements are these indices, namely

$$\Gamma = \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_n \\ \gamma_{n+1} & \gamma_{n+2} & \cdots & \gamma_{2n} \\ \vdots & \vdots & & \vdots \\ \gamma_{(m-1)n+1} & \gamma_{(m-1)n+2} & \cdots & \gamma_{mn} \end{bmatrix}.$$

We take as the product $\gamma_{n_1} \cdot \gamma_{n_2}$ of any two elements γ_{n_1} and γ_{n_2} of the matrix Γ the ordered pair $(\gamma_{n_1}, \gamma_{n_2})$ and form the direct product of Γ with itself k -times on the basis of this notion for the "product" of the elements of Γ . Thus, for example,

$$\Gamma \otimes \Gamma = [\gamma_{n_1}]_{n_1} \otimes \Gamma = [\gamma_{n_1} \cdot \Gamma]_{n_1} = [[\gamma_{n_1} \cdot \gamma_{n_2}]_{n_2}]_{n_1} = [[(\gamma_{n_1}, \gamma_{n_2})]_{n_2}]_{n_1}.$$

Keeping these remarks in mind we now present

Theorem 2.6.2. Provided the required k -th order scalar partial derivatives exist and are continuous, the k -th order Dwyer-Macphail matrix derivative with respect to the index $\pi = (r, s)$ of the matrix function $Y = F(X) \in \Phi_{(p, q)}$ of the matrix variable $X \in \Phi_{(m, n)}$ is given by the $mp \times mq$ -matrix

$$\frac{\partial^k \langle Y \rangle_\pi}{\partial X^k} = \left[\frac{\partial^k \langle Y \rangle_{(\pi, \xi_2, \dots, \xi_k)}}{\partial X^k} \right]$$

where the indices $\xi_2 = \gamma_{m_1}, \dots, \xi_k = \gamma_{m_{k-1}}$ are chosen by identifying the matrix block $\frac{\partial^k \langle Y \rangle_{(\pi, \xi_2, \dots, \xi_k)}}{\partial X^k}$ of $\frac{\partial^k \langle Y \rangle}{\partial X^k}$ with the identically positioned element $\gamma_{m_1} \cdot \gamma_{m_2} \cdots \gamma_{m_{k-1}}$ in the direct product of the matrix Γ

with itself (k-1)-times and taking the factors of this product
 $(\gamma_{m_1}, \gamma_{m_2}, \dots, \gamma_{m_{k-1}})$ in the order in which they appear as the values for
 $\xi_2, \dots, \xi_k$ in the same order in the indexing set for the corresponding
matrix block $\frac{\partial^k \langle Y \rangle_{(\pi, \xi_2, \dots, \xi_k)}}{\partial X^k}$ of $\frac{\partial^k \langle Y \rangle_{\pi}}{\partial X^k}$.

Proof. From the Definition 2.6.2 of the k-th order Dwyer-Macphail matrix derivative we find that

$$\begin{aligned} \frac{\partial^k \langle Y \rangle_{\pi}}{\partial X^k} &= \left[\frac{\partial^{k-1}}{\partial \langle X \rangle_{\eta_{k-1}} \partial \langle X \rangle_{\eta_{k-2}} \dots \partial \langle X \rangle_{\eta_1}} \left(\frac{\partial \langle Y \rangle_{\pi}}{\partial X} \right) \right]_{\{\eta_j\}} \\ &= \left[\frac{\partial}{\partial X} \left(\frac{\partial^{k-1} \langle Y \rangle_{\pi}}{\partial \langle X \rangle_{\eta_{k-1}} \partial \langle X \rangle_{\eta_{k-2}} \dots \partial \langle X \rangle_{\eta_1}} \right) \right]_{\{\eta_j\}} \\ &= \left[\frac{\partial^k \langle Y \rangle_{(\pi, \xi_2, \dots, \xi_k)}}{\partial X^k} \right]_{\{\xi_{j+1} = \eta_j\}}, \end{aligned}$$

and thus the proof of this theorem is simple and direct.

To provide an illustration of these notions let us find the higher order Dwyer-Macphail matrix derivatives of the matrix function

$Y = XAX^T BX = F(X)$ where $A, B, X \in \Phi(2, 2)$ which we examined in some detail in Section 2.5. For this function we find that,

$$\frac{\partial \langle Y \rangle_{\pi}}{\partial X} = K_{\pi} X^T B^T X A^T + B X K_{\pi}^T X A + B^T X A^T X^T K_{\pi};$$

and

$$\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} = \begin{bmatrix} \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_1)}}{\partial X^2} & \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_2)}}{\partial X^2} \\ \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_3)}}{\partial X^2} & \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_4)}}{\partial X^2} \end{bmatrix}$$

where for $k = 1, 2, 3, 4$,

$$\begin{aligned} \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_k)}}{\partial X^2} &= B X K_\pi^T J_{\gamma_k} A + B^T X A^T J_{\gamma_k}^T K_\pi + K_\pi X^T B^T J_{\gamma_k} A^T \\ &+ B^T J_{\gamma_k} A^T X^T K_\pi + K_\pi J_{\gamma_k}^T B^T X A^T + B J_{\gamma_k} K_\pi^T X A. \end{aligned}$$

Next, to determine $\frac{\partial^3 \langle Y \rangle_\pi}{\partial X^3}$, we first compute the direct product of the matrix

$$\Gamma = \begin{bmatrix} \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_4 \end{bmatrix}$$

with itself two times where $\gamma_1 = (1,1)$, $\gamma_2 = (1,2)$, $\gamma_3 = (2,1)$ and $\gamma_4 = (2,2)$. We obtain, following the various definitions required,

$$\Gamma \otimes \Gamma = \begin{bmatrix} (\gamma_1, \gamma_1) & (\gamma_1, \gamma_2) & (\gamma_2, \gamma_1) & (\gamma_2, \gamma_2) \\ (\gamma_1, \gamma_3) & (\gamma_1, \gamma_4) & (\gamma_2, \gamma_3) & (\gamma_2, \gamma_4) \\ (\gamma_3, \gamma_1) & (\gamma_3, \gamma_2) & (\gamma_4, \gamma_1) & (\gamma_4, \gamma_2) \\ (\gamma_3, \gamma_3) & (\gamma_3, \gamma_4) & (\gamma_4, \gamma_3) & (\gamma_4, \gamma_4) \end{bmatrix}$$

and; therefore,

$$\frac{\partial^3 \langle Y \rangle_{\pi}}{\partial X^3} = \begin{bmatrix} \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_1, \gamma_1)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_1, \gamma_2)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_2, \gamma_1)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_2, \gamma_2)}}{\partial X^3} \\ \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_1, \gamma_3)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_1, \gamma_4)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_2, \gamma_3)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_2, \gamma_4)}}{\partial X^3} \\ \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_3, \gamma_1)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_3, \gamma_2)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_4, \gamma_1)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_4, \gamma_2)}}{\partial X^3} \\ \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_3, \gamma_3)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_3, \gamma_4)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_4, \gamma_3)}}{\partial X^3} & \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_4, \gamma_4)}}{\partial X^3} \end{bmatrix}$$

where, for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$,

$$\begin{aligned} \frac{\partial^3 \langle Y \rangle_{(\pi, \gamma_i, \gamma_j)}}{\partial X^3} &= B^T J_{\gamma_j} A^T J_{\gamma_i}^T K_{\pi} + B J_{\gamma_j} K_{\pi}^T J_{\gamma_i} A + B^T J_{\gamma_i} A^T J_{\gamma_j}^T K_{\pi} \\ &+ K_{\pi} J_{\gamma_j}^T B^T J_{\gamma_i} A^T + B J_{\gamma_i} K_{\pi}^T J_{\gamma_j} A + K_{\pi} J_{\gamma_i}^T B^T J_{\gamma_j} A^T . \end{aligned}$$

Finally, for $k \geq 4$, we find that

$$\frac{\partial^k \langle Y \rangle_{\pi}}{\partial X^k} = 0 \cdot I_{2^k \times 2^k} .$$

This, then, completes the derivation of the Dwyer-Macphail matrix derivatives of all orders for the given matrix function $Y = XAX^T BX = F(X)$.

We now go on to show that these Dwyer-Macphail matrix derivatives play an important role in the question of whether or not a matrix function of a matrix variable can be expanded in an (infinite) matrix series analogous in form to the Taylor's series expansion of a scalar function $y = f(\lambda)$ of a scalar variable, namely

$$f(\lambda) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(\lambda_0) (\lambda - \lambda_0)^k.$$

In other words is it possible to replace the $f^{(k)}(\lambda_0)$ by the Dwyer-Macphail matrix derivatives $\frac{\partial^k \langle F \rangle}{\partial X^k}$ in this formula when instead of a scalar function of a scalar variable we consider a matrix function $Y = F(X)$ of a matrix variable X ?

Of the previous attempts that have been made in this direction, Miss Parel in [114] succeeded in giving a matrix development for series expansions of scalar functions of vector variables. In order to accomplish this she defined certain types of higher order matrix derivatives for scalar functions of vector variables which are different than the higher order matrix derivatives that have been introduced here. More specifically consider a scalar $y = f(x)$ of a column matrix (vector) $x \in \Phi_{(m,1)}$. For this function, of course, the first order Dwyer-Macphail matrix derivative $\partial f / \partial x$ is also a column matrix (vector) belonging to $\Phi_{(m,1)}$. Miss Parel then defined a second order type of matrix derivative by

$$\left(\frac{\partial}{\partial x} \right) \left(\frac{\partial f}{\partial x} \right)^T = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{m \times m}$$

which can be interpreted as that $m \times m$ -matrix obtained by (matrix) multiplication of the column matrix operator $\left(\frac{\partial}{\partial x}\right) \epsilon_{(m,1)}$ of scalar partial differentiation operators formed from the components of x by the row matrix $\left(\frac{\partial f}{\partial x}\right)^T \epsilon_{(1,m)}$ of scalar partial derivatives of f with respect to the components of x . She then introduced a mixed type of matrix derivative, $\frac{\partial}{\partial \langle x \rangle_i} \left(\frac{\partial f}{\partial x}\right)^T$ for $i = 1, 2, \dots, m$, and noting that

$$\left(\frac{\partial}{\partial x}\right) \left(\frac{\partial f}{\partial x}\right)^T = \begin{bmatrix} \frac{\partial}{\partial \langle x \rangle_1} & \left(\frac{\partial f}{\partial x}\right)^T \\ \frac{\partial}{\partial \langle x \rangle_2} & \left(\frac{\partial f}{\partial x}\right)^T \\ \vdots & \vdots \\ \frac{\partial}{\partial \langle x \rangle_m} & \left(\frac{\partial f}{\partial x}\right)^T \end{bmatrix},$$

she introduced a new rule for deriving $\left(\frac{\partial}{\partial x}\right) \left(\frac{\partial f}{\partial x}\right)^T$ from the $\frac{\partial}{\partial \langle x \rangle_i} \left(\frac{\partial f}{\partial x}\right)^T$.

Finally she defined other mixed types of higher order matrix derivatives of the form

$$\begin{aligned} & \left(\frac{\partial^s}{\partial \langle x \rangle_{i_1} \partial \langle x \rangle_{i_2} \dots \partial \langle x \rangle_{i_s}} \right) \left(\frac{\partial}{\partial x}\right) \left(\frac{\partial f}{\partial x}\right)^T \\ &= \left[\frac{\partial^s}{\partial \langle x \rangle_{i_1} \partial \langle x \rangle_{i_2} \dots \partial \langle x \rangle_{i_s}} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right) \right]_{(i,j)} \end{aligned}$$

and then developed a series expansion for scalar functions of vector variables which featured these types of higher order matrix derivatives.

As previously noted, in her approach she gave new rules in order to effectively derive the higher order matrix derivatives of the types that she had introduced and in terms of which she presented this matrix development for series expansions of scalar functions of vector variables. The application of these new rules for deriving higher order matrix derivatives of the type used by Miss Parel, while reasonably convenient for obtaining second order matrix derivatives of this type, can become quite complicated for third and higher order matrix derivatives of the type that she studied. Thus, in effect, for many scalar functions of vector variables, one is frequently limited to an explicit development in matrix notation of only second degree series expansions when her types of higher order matrix derivatives are used.

Further, neither the higher order matrix derivatives introduced by Miss Parel nor the rules that she presented for their derivation appear to be readily adaptable to scalar functions of a matrix variable or matrix functions of a matrix variable. Thus her higher order matrix derivatives do not possess the same degree of simplicity nor the same degree of generality as do the higher order Dwyer-Macphail matrix derivatives introduced for the first time in this work.

Returning to our own development for an (infinite) matrix series expansion of a matrix function of a matrix variable which features its Dwyer-Macphail matrix derivatives, the following theorem expresses a key

observation upon which we build such a series expansion:

Theorem 2.6.3. If the scalar function $y = f(X)$ of the matrix variable $X \in \Phi_{(m,n)}$ has Dwyer-Macphail matrix derivatives of all orders, then,
for each positive integer k ,

$$\text{Tr} \left(\frac{\partial^k f}{\partial X^k} (H^T)^k \right) = \sum_{k_1=0}^k \sum_{k_2=0}^k \cdots \sum_{k_{mn}=0}^k \frac{\partial^k f}{\partial x_{\gamma_1}^{k_1} \partial x_{\gamma_2}^{k_2} \cdots \partial x_{\gamma_{mn}}^{k_{mn}}} h_{\gamma_1}^{k_1} h_{\gamma_2}^{k_2} \cdots h_{\gamma_{mn}}^{k_{mn}} \\ (k_1 + k_2 + \cdots + k_{mn} = k)$$

where $H \in \Phi_{(m,n)}$ is any matrix whatsoever.

Proof. Clearly this theorem is true for the case $k = 1$ since we have in this case

$$\begin{aligned} \frac{\partial f}{\partial X} H^T &= \left[\frac{\partial f}{\partial \langle X \rangle_{ij}} \right] [h_{ij}]^T \\ &= \left[\frac{\partial f}{\partial \langle X \rangle_{ij}} \right] [h_{ji}] \\ &= \left[\sum_{q=1}^n \frac{\partial f}{\partial \langle X \rangle_{iq}} h_{jq} \right]_{(i,j)} \end{aligned}$$

Hence we see that,

$$\text{Tr} \left(\frac{\partial f}{\partial X} H^T \right) = \sum_{p=1}^m \sum_{q=1}^n \frac{\partial f}{\partial \langle X \rangle_{pq}} h_{pq} \quad .$$

Next, suppose we assume that the theorem is true for a particular k -value, $k > 1$. Now, we first note that

$$\begin{aligned}
\frac{\partial^{k+1} f}{\partial X^{k+1}} (H^T)^{k+1} &= \left(\frac{\partial}{\partial X} \otimes \frac{\partial^k f}{\partial X^k} \right) (H^T \otimes (H^T)^k) \\
&= \left[\frac{\partial}{\partial \langle X \rangle_{ij}} \left(\frac{\partial^k f}{\partial X^k} \right) \right] \cdot [h_{ji} (H^T)^k] \\
&= \left[\sum_{q=1}^n \frac{\partial}{\partial \langle X \rangle_{iq}} \left(\frac{\partial^k f}{\partial X^k} (H^T)^k \right) h_{jq} \right]_{(i,j)}.
\end{aligned}$$

From this expression, we see that

$$\text{Tr} \left(\frac{\partial^{k+1} f}{\partial X^{k+1}} (H^T)^{k+1} \right) = \sum_{p=1}^m \sum_{q=1}^n \frac{\partial}{\partial \langle X \rangle_{pq}} \text{Tr} \left(\frac{\partial^k f}{\partial X^k} (H^T)^k \right) h_{pq}$$

But, since our induction hypothesis is that

$$\text{Tr} \left(\frac{\partial^k f}{\partial X^k} (H^T)^k \right) = \sum_{k_1=0}^k \sum_{k_2=0}^k \cdots \sum_{k_{mn}=0}^k \frac{\partial^k f}{\partial X_{\gamma_1}^{k_1} \partial X_{\gamma_2}^{k_2} \cdots \partial X_{\gamma_{mn}}^{k_{mn}}} h_{\gamma_1}^{k_1} h_{\gamma_2}^{k_2} \cdots h_{\gamma_{mn}}^{k_{mn}},$$

($k_1 + k_2 + \cdots + k_{mn} = k$)

we see, upon substitution of this expression for $\text{Tr} \left(\frac{\partial^k f}{\partial X^k} (H^T)^k \right)$ into the given expression for $\text{Tr} \left(\frac{\partial^{k+1} f}{\partial X^{k+1}} (H^T)^{k+1} \right)$ and making notational adjustments that this theorem also holds for the case $k+1$. Consequently, through mathematical induction, we conclude that this theorem holds for all k -values.

We remark here that by adopting the convention that for $k = 0$ the symbol $(H^T)^0$ is to be interpreted as J_{11}^T , since by previous agreement we have written $f = \frac{\partial^0}{\partial X^0} f$, we find under these interpretations that

$$\text{Tr} \left(\frac{\partial^0 f}{\partial X^0} (H^T)^0 \right) = \text{Tr}(f J_{11}^T) = f \text{Tr} J_{11}^T = f;$$

and thus Theorem 2.6.3 can be extended in this manner to all non-negative integers and this is understood to be the case in the subsequent developments.

The preceding results have brought us close to our goal of giving a matrix series development for a matrix function of a matrix variable featuring its Dwyer-Macphail matrix derivatives. Thus let us first consider a scalar function $y = f(X)$ of a matrix variable $X \in \Phi_{(m,n)}$. In this connection we write, where $\bar{X} \in \Phi_{(m,n)}$ is a fixed but arbitrary matrix and $H = X - \bar{X} = [x_{ij} - \bar{x}_{ij}]_{(i,j)} = [h_{ij}]_{(i,j)}$,

$$X = \begin{bmatrix} x_{\gamma_1} & x_{\gamma_2} & \dots & x_{\gamma_n} \\ x_{\gamma_{n+1}} & x_{\gamma_{n+2}} & \dots & x_{\gamma_{2n}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{\gamma_{(m-1)n+1}} & x_{\gamma_{(m-1)n+2}} & \dots & x_{\gamma_{mn}} \end{bmatrix},$$

$$\bar{X} = \begin{bmatrix} \bar{x}_{\gamma_1} & \bar{x}_{\gamma_2} & \dots & \bar{x}_{\gamma_n} \\ \bar{x}_{\gamma_{n+1}} & \bar{x}_{\gamma_{n+2}} & \dots & \bar{x}_{\gamma_{2n}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \bar{x}_{\gamma_{(m-1)n+1}} & \bar{x}_{\gamma_{(m-1)n+2}} & \dots & \bar{x}_{\gamma_{mn}} \end{bmatrix},$$

and

$$H = \begin{bmatrix} h_{\gamma_1} & h_{\gamma_2} & \dots & h_{\gamma_n} \\ h_{\gamma_{n+1}} & h_{\gamma_{n+2}} & \dots & h_{\gamma_{2n}} \\ \vdots & \vdots & & \vdots \\ h_{\gamma_{(m-1)n+1}} & h_{\gamma_{(m-1)n+2}} & \dots & h_{\gamma_{mn}} \end{bmatrix}.$$

Suppose that the representation of $y = f(X)$ as a scalar function of several scalar variables $y = f(x)$ where $x = (x_{\gamma_1}, x_{\gamma_2}, \dots, x_{\gamma_{mn}})$ can be expanded in a Taylor's series about the point $\bar{x} = (\bar{x}_{\gamma_1}, \bar{x}_{\gamma_2}, \dots, \bar{x}_{\gamma_{mn}})$; in other words,

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\sum_{k_1=0}^k \sum_{k_2=0}^k \dots \sum_{k_{mn}=0}^k \frac{\partial^k f}{\partial x_{\gamma_1}^{k_1} \partial x_{\gamma_2}^{k_2} \dots \partial x_{\gamma_{mn}}^{k_{mn}}} h_{\gamma_1}^{k_1} h_{\gamma_2}^{k_2} \dots h_{\gamma_{mn}}^{k_{mn}} \right) .$$

($k_1+k_2+\dots+k_{mn}=k$)

Under these circumstances we say that the scalar function $y = f(X)$ of the matrix variable $X \in \Phi_{(m,n)}$ has a "Dwyer-Macphail matrix derivative series expansion about $X = \bar{X}$ " given by the infinite series

$$f(X) = \sum_{k=0}^{\infty} \frac{1}{k!} \text{Tr} \left(\left[\frac{\partial^k f}{\partial X^k} \right]_{X=\bar{X}} (H^T)^k \right) ;$$

and this series expansion for a scalar function of a matrix variable featuring its Dwyer-Macphail matrix derivatives is justified in light of Theorem 2.6.3.

We note, moreover, for a row matrix (vector) $x \in \Phi_{(1,n)}$ that we obtain the Dwyer-Macphail matrix derivative series expansion in the form

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} \left[\frac{\partial^k f}{\partial x^k} \right]_{x=\bar{x}} (h^T)^k$$

where $h = x - \bar{x}$ and $\bar{x}$ are elements of $\Phi_{(1,n)}$ too. Thus the Dwyer-Macphail matrix derivative series expansion of a scalar function of a vector variable is directly analogous in form to the Taylor's series expansion of a scalar function of a scalar variable.

Finally, let us consider a matrix function of a matrix variable from $\Phi_{(m,n)}$ to $\Phi_{(p,q)}$, say

$$Y = [f_{rs}(X)]_{(r,s)} = F(X) ;$$

as before consider $\bar{X} \in \Phi_{(m,n)}$ and $H = X - \bar{X}$. If the representation $y_{rs} = f_{rs}(x)$ where $x \in \Phi_{(1, mn)}$ of each element function $y_{rs} = f_{rs}(X)$ of the matrix function $Y = F(X)$ has a Taylor's series expansion about $x = \bar{x}$ we would have for each pair (r,s) with $1 \leq r \leq p$ and $1 \leq s \leq q$.

$$f_{rs}(X) = \sum_{k=0}^{\infty} \frac{1}{k!} \text{Tr} \left(\frac{\partial^k \langle Y \rangle_{rs}}{\partial X^k} \right)_{X=\bar{X}} (H^T)^k ;$$

and then under such circumstances we say as in

Definition 2.6.3. The matrix function $Y = F(X) \in \Phi_{(p,q)}$ of the matrix variable $X \in \Phi_{(m,n)}$ has a Dwyer-Macphail matrix derivative series expansion about $X = \bar{X}$ given by the (infinite) matrix series

$$F(X) = \sum_{k=0}^{\infty} \frac{1}{k!} \left[\text{Tr} \left(\frac{\partial^k \langle F \rangle_{rs}}{\partial X^k} \right)_{X=\bar{X}} (H^T)^k \right]_{(r,s)} .$$

In this manner we have obtained an extension of Taylor's theorem for scalar functions of a scalar variable to matrix functions of a matrix variable which features the Dwyer-Macphail matrix derivatives of its element functions. Moreover, the same expansion holds equally as well, if instead of a matrix function of a matrix variable, we consider either a scalar function of a matrix variable or a scalar function of a vector variable or simply a scalar function of a scalar variable.

It might be instructive at this point to give an illustration of the Dwyer-Macphail matrix derivative series expansion for a specific matrix function. Let us consider, therefore, the matrix function of a matrix variable

$$Y = XAX^T = F(X)$$

where $A, X \in \Phi_{(2,2)}$ and where we write

$$X = \begin{bmatrix} x_{\gamma_1} & x_{\gamma_2} \\ x_{\gamma_3} & x_{\gamma_4} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} y_{\pi_1} & y_{\pi_2} \\ y_{\pi_3} & y_{\pi_4} \end{bmatrix} .$$

Now applying the appropriate theorems from Section 2.5 we find that for this matrix function of a matrix variable,

$$\frac{\partial Y}{\partial \langle X \rangle_\gamma} = J_\gamma A X^T + X A J_\gamma^T ;$$

$$\frac{\partial}{\partial \langle X \rangle_\gamma} \left(\frac{\partial \langle Y \rangle_\pi}{\partial X} \right) = K_{\pi \gamma} J_\gamma A^T + K_{\pi \gamma}^T J_\gamma A ;$$

$$\frac{\partial^2 \langle Y \rangle_{(\pi, \gamma)}}{\partial X^2} = K_{\pi \gamma}^T J_\gamma A + K_{\pi \gamma} J_\gamma A^T ;$$

and, for $k \geq 3$,

$$\frac{\partial^k \langle Y \rangle_{(\pi, \xi_2, \dots, \xi_k)}}{\partial X^k} = 0 \cdot I_{2 \times 2} .$$

In turn we find that

$$\frac{\partial \langle Y \rangle_\pi}{\partial X} = K_\pi X A^T + K_\pi^T X A ;$$

and

$$\begin{aligned} \frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} &= \begin{bmatrix} \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_1)}}{\partial X^2} & \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_2)}}{\partial X^2} \\ \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_3)}}{\partial X^2} & \frac{\partial^2 \langle Y \rangle_{(\pi, \gamma_4)}}{\partial X^2} \end{bmatrix} \\ &= \begin{bmatrix} K_{\pi \gamma_1}^T K_{\gamma_1} A & K_{\pi \gamma_2}^T K_{\gamma_2} A \\ K_{\pi \gamma_3}^T K_{\gamma_3} A & K_{\pi \gamma_4}^T K_{\gamma_4} A \end{bmatrix} \\ &+ \begin{bmatrix} K_{\pi \gamma_1} K_{\gamma_1} A^T & K_{\pi \gamma_2} K_{\gamma_2} A^T \\ K_{\pi \gamma_3} K_{\gamma_3} A^T & K_{\pi \gamma_4} K_{\gamma_4} A^T \end{bmatrix} ; \end{aligned}$$

and, finally, for $k \geq 3$,

$$\frac{\partial^k \langle Y \rangle_\pi}{\partial X^k} = 0 \cdot I_{2^k \times 2^k}.$$

Using these results we now seek the individual terms of the series expansion for $Y = XAX^T$ featuring its Dwyer-Macphail matrix derivatives.

Thus we note first that

$$\begin{aligned} \text{Tr} \left(\left[\frac{\partial \langle Y \rangle_\pi}{\partial X} \right]_{X=\bar{X}} (X-\bar{X})^T \right) &= \text{Tr}((K_\pi \bar{X} A^T + K_\pi^T \bar{X} A)(X-\bar{X})^T) \\ &= \text{Tr}(K_\pi \bar{X} A^T (X-\bar{X})^T) + \text{Tr}(K_\pi^T \bar{X} A (X-\bar{X})^T) \\ &= \text{Tr}(K_\pi^T (X-\bar{X}) A \bar{X}^T) + \text{Tr}(\bar{X} A (X-\bar{X})^T K_\pi^T) \\ &= \text{Tr}((X-\bar{X}) A \bar{X}^T K_\pi^T) + \text{Tr}(\bar{X} A (X-\bar{X})^T K_\pi^T) \\ &= \langle (X-\bar{X}) A \bar{X}^T \rangle_\pi + \langle \bar{X} A (X-\bar{X})^T \rangle_\pi. \end{aligned}$$

Therefore we find that

$$\begin{aligned} \left[\text{Tr} \left(\left[\frac{\partial \langle Y \rangle_\pi}{\partial X} \right]_{X=\bar{X}} (X-\bar{X})^T \right) \right]_\pi &= \left[\langle (X-\bar{X}) A \bar{X}^T \rangle_\pi \right]_\pi + \left[\langle \bar{X} A (X-\bar{X})^T \rangle_\pi \right]_\pi \\ &= (X-\bar{X}) A \bar{X}^T + \bar{X} A (X-\bar{X})^T. \end{aligned}$$

Next we write

$$((X-\bar{X})^T)^2 = (H^T)^2 = \begin{bmatrix} h_{11} H^T & h_{21} H^T \\ h_{12} H^T & h_{22} H^T \end{bmatrix}$$

and find that

$$\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} \cdot (H^T)^2 =$$

$$\begin{aligned} & \begin{bmatrix} K_\pi^T K_{\gamma_1} A h_{11} H^T + K_\pi^T K_{\gamma_2} A h_{12} H^T & K_\pi^T K_{\gamma_1} A h_{21} H^T + K_\pi^T K_{\gamma_2} A h_{22} H^T \\ K_\pi^T K_{\gamma_3} A h_{11} H^T + K_\pi^T K_{\gamma_4} A h_{12} H^T & K_\pi^T K_{\gamma_3} A h_{21} H^T + K_\pi^T K_{\gamma_4} A h_{22} H^T \end{bmatrix} \\ & + \begin{bmatrix} K_\pi K_{\gamma_1} A^T h_{11} H^T + K_\pi K_{\gamma_2} A^T h_{12} H^T & K_\pi K_{\gamma_1} A^T h_{21} H^T + K_\pi K_{\gamma_2} A^T h_{22} H^T \\ K_\pi K_{\gamma_3} A^T h_{11} H^T + K_\pi K_{\gamma_4} A^T h_{12} H^T & K_\pi K_{\gamma_3} A^T h_{21} H^T + K_\pi K_{\gamma_4} A^T h_{22} H^T \end{bmatrix} \end{aligned}$$

From this expression for $\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} \cdot (H^T)^2$ we obtain the result that

$$\begin{aligned} \text{Tr} \left(\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} \right)_{X=\bar{X}} (X^T - \bar{X}^T)^2 &= \text{Tr}(K_\pi^T K_{\gamma_1} A h_{11} H^T + K_\pi^T K_{\gamma_2} A h_{12} H^T) \\ &+ \text{Tr}(K_\pi^T K_{\gamma_3} A h_{21} H^T + K_\pi^T K_{\gamma_4} A h_{22} H^T) \\ &+ \text{Tr}(K_\pi K_{\gamma_1} A^T h_{11} H^T + K_\pi K_{\gamma_2} A^T h_{12} H^T) \\ &+ \text{Tr}(K_\pi K_{\gamma_3} A^T h_{21} H^T + K_\pi K_{\gamma_4} A^T h_{22} H^T) \end{aligned}$$

Once again using the properties of the Tr-operator established in Theorem

2.4.1 we see that

$$\begin{aligned} \text{Tr} \left(\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} \right)_{X=\bar{X}} (X^T - \bar{X}^T)^2 &= \text{Tr}(h_{\gamma_1} K_{\gamma_1} A H^T K_\pi^T) + \text{Tr}(h_{\gamma_2} K_{\gamma_2} A H^T K_\pi^T) \\ &+ \text{Tr}(h_{\gamma_3} K_{\gamma_3} A H^T K_\pi^T) + \text{Tr}(h_{\gamma_4} K_{\gamma_4} A H^T K_\pi^T) \\ &+ \text{Tr}(h_{\gamma_1} H A K_{\gamma_1}^T K_\pi^T) + \text{Tr}(h_{\gamma_2} H A K_{\gamma_2}^T K_\pi^T) \\ &+ \text{Tr}(h_{\gamma_3} H A K_{\gamma_3}^T K_\pi^T) + \text{Tr}(h_{\gamma_4} H A K_{\gamma_4}^T K_\pi^T) ; \end{aligned}$$

or, equivalently,

$$\text{Tr} \left(\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} \right)_{X=\bar{X}} (X^T - \bar{X}^T)^2 = \sum_{k=1}^4 \left(\langle h_{\gamma_k} K_{\gamma_k} A H^T \rangle_\pi + \langle h_{\gamma_k} H A K_{\gamma_k}^T \rangle_\pi \right).$$

We see, therefore, that the 2-nd degree term of the Dwyer-Macphail matrix derivative series expansion of $Y = XAX^T$ about $X = \bar{X}$ is given by the expression

$$\begin{aligned} \left[\text{Tr} \left(\frac{\partial^2 \langle Y \rangle_\pi}{\partial X^2} \right)_{X=\bar{X}} (X^T - \bar{X}^T)^2 \right]_\pi &= \sum_{k=1}^4 \left[\langle h_{\gamma_k} K_{\gamma_k} A H^T \rangle_\pi + \langle h_{\gamma_k} H A K_{\gamma_k}^T \rangle_\pi \right]_\pi \\ &= \sum_{k=1}^4 (h_{\gamma_k} K_{\gamma_k} A H^T + h_{\gamma_k} H A K_{\gamma_k}^T) \\ &= \sum_{k=1}^4 (h_{\gamma_k} K_{\gamma_k} A H^T) + \sum_{k=1}^4 (H A h_{\gamma_k} K_{\gamma_k}^T) \\ &= \left(\sum_{k=1}^4 h_{\gamma_k} K_{\gamma_k} \right) A H^T + H A \left(\sum_{k=1}^4 h_{\gamma_k} K_{\gamma_k}^T \right) \\ &= H A H^T + H A H^T \\ &= 2(X - \bar{X}) A (X - \bar{X})^T. \end{aligned}$$

Now, since for $k \geq 3$,

$$\frac{\partial^k \langle Y \rangle_\pi}{\partial X^k} = 0 \cdot I_{2^k \times 2^k},$$

3rd and higher order terms of the matrix series expansion for $Y = XAX^T$ are merely the zero matrix $0 \in \Phi_{(2,2)}$. Therefore the matrix series expansion of the matrix function $Y = XAX^T = F(X) \in \Phi_{(2,2)}$ of the matrix variable $X \in \Phi_{(2,2)}$ about the matrix $X = \bar{X}$ obtained by using its Dwyer-Macphail ma-

trix derivatives is given by the expression

$$XAX^T = \bar{X}A\bar{X}^T + \bar{X}A(X-\bar{X})^T + (\bar{X}-X)A\bar{X}^T + (X-\bar{X})A(X-\bar{X})^T .$$

When $A = A^T$,

$$XA(X-\bar{X})^T = ((X-\bar{X})A\bar{X}^T)^T ,$$

and we may write this matrix series expansion in the form

$$XAX^T = \bar{X}A\bar{X}^T + \bar{X}A(X-\bar{X})^T + (\bar{X}A(X-\bar{X})^T)^T + (X-\bar{X})A(X-\bar{X})^T .$$

We note the analogy between this formula for the Dwyer-Macphail matrix derivative series expansion of the matrix function $Y = XAX^T = F(X) \in \Phi_{(2,2)}$ about $X = \bar{X}$ and the Taylor's series expansion of the scalar function $y = ax^2$ of the scalar variable x about $x = \bar{x}$, namely

$$ax^2 = a\bar{x}^2 + 2a\bar{x}(x-\bar{x}) + a(x-\bar{x})^2 .$$

With the discussion of this example completed, we bring to a close these direct extensions of the calculus of Dwyer-Macphail matrix derivatives. We return to these ideas in later chapters when we undertake a development of the problem of estimating the unknown parameters of complex stochastic structures representing linear structural relationships having both "errors-in-variables" and "errors-in-equations." First, however, before applying the Dwyer-Macphail matrix derivative calculus to these important estimation problems in statistics, we develop some interesting and basic connections between the Dwyer-Macphail matrix derivative and the **Frechet-Gateaux** matrix differential in the next chapter.

CHAPTER 3

DWYER-MACPHAIL MATRIX DERIVATIVES AND FRECHET-GATEAUX MATRIX DIFFERENTIALS

3.1. INTRODUCTION

Besides the use of the ordinary differential calculus for scalar functions of several scalar variables or the more advanced techniques of Dwyer-Macphail matrix differentiation as applied to matrix functions of a matrix variable, one occasionally finds in the literature another method of maximization or minimization involving the use of differentials of scalar or matrix functions of vector or matrix variables. One such illustration involving the use of differentials of a matrix function of a matrix variable is provided in [21] where Chipman and Rao , using matrix differentials, show that when A is an rxn -matrix, X is a fixed $n \times k$ -matrix of rank k , and ψ is a fixed rxk -matrix, then the diagonal elements of AA^T reach their simultaneous minima subject to the constraints $AX = \psi$ when $A = \psi(X^T X)^{-1} X^T$.

Chipman and Rao did not indicate the conditions under which such matrix differential techniques can be applied to yield valid mathematical derivations nor did they suggest the possibility that these matrix differential methods might be related to other well established techniques of matrix differentiation.

Actually these matrix differential methods can be given a rigorous mathematical formulation, and they are intimately connected with the

Dwyer-Macphail matrix derivative methods. In this chapter we undertake a more complete investigation of these issues and in the course of resolving them we discover new extensions of the Dwyer-Macphail matrix derivatives in the form of certain basic relationships which exist between them and the Gateaux and Frechet differentials of matrix functions of matrix variables.

The fundamental notions of differentials, polynomials and power series as they apply to abstract functions were introduced in [36,37] by M. Frechet around 1909. A few years later R. Gateaux in [39,40] applied the ideas of Frechet to certain concrete problems and his work which was published posthumously provided a basis for many of the subsequent developments in this area of functional analysis.

Since these initial efforts, the theory of Frechet and Gateaux differentials of abstract functions has been greatly expanded and refined by such mathematicians as A. D. Michal [103], I. E. Highberg [59] and R. S. Martin [102], who were students of Michal, and by L. M. Graves [50] and A. E. Taylor [135]. Recently in [94], R. M. Lorch developed a somewhat different notion of differentiability for functions with domain and range in a commutative Banach algebra.

In several of the following sections of this chapter we have summarized in their present form parts of the theory of Frechet and Gateaux differentials for functions from and to real or complex Banach spaces or Banach algebras. This summary deals mainly with those parts of the theory which are especially significant to our work involving matrix deriv-

atives and matrix differentials.

The somewhat explicit and detailed presentation accorded these results here has been made primarily to establish an orientation for our main results on matrix derivatives and differentials presented in Section 3.5. But also it is hoped that by making our review of these general notions at this level, the reading of this thesis can be facilitated by reducing the number of cross-references to other works that would otherwise be necessary in order to follow some of the proofs for the new developments relating to derivatives and differentials of matrix functions of matrix variables.

Under certain regularity conditions the Frechet differential and the Gateaux differential of an abstract function co-exist and are co-equal. Under these circumstances we refer merely to the Frechet-Gateaux differential of the abstract function. It turns out in the case of a function from and to matrix Banach spaces or algebras over the real or complex fields that Frechet-Gateaux matrix differentials are uniquely specified in terms of the Dwyer-Macphail matrix derivatives of the function. In other words, there exists an abstract homogeneous matrix polynomial completely defined in terms of the Dwyer-Macphail matrix derivatives which is in fact just the Frechet-Gateaux matrix differential of a given matrix function of a matrix variable.

Other notions of differentials for abstract functions, as well as matrix derivatives for matrix functions, have been proposed by various authors and their properties explored. We have already mentioned the

notion of Lorch differentiability for abstract functions whose domain and range are a commutative Banach algebra. Along a different approach, however, F. Hausdorff in [57] defined a notion of differential for functions on and to an associative linear algebra with a unit element over the real or complex field. The Hausdorff differential df of a function $z = f(x)$ from and to an associative linear algebra A with unit element over the real or complex field Φ is an additive homogeneous function of the differential dx of the form

$$df = \sum_{i=1}^n u_i dx v_i$$

where u_i and v_i are elements of the algebra A which depend only on x . Based on this notion of differential, Hausdorff proceeded to develop a theory of analytic functions for associative hyper-complex systems having a unit element.

Hausdorff's work was published in 1900; in 1957, R. Rinehart in [123] defined a notion of matrix derivative for matrix function from and to the matrix algebra $\Phi_{(k,k)}$ of square matrices of order k . The Rinehart matrix derivative $Z = F^*(X)$ of a matrix function $Y = F(X)$ on and to $\Phi_{(k,k)}$ is defined as

$$F^*(X) = \lim_{m \rightarrow \infty} \sum_{i=1}^n U_i^{(m)} V_i^{(m)}$$

where $U_i^{(m)}$ and $V_i^{(m)}$ are matrices belonging to $\Phi_{(k,k)}$ for which

$$F(X+H_m) - F(X) = \sum_{i=1}^n U_i^{(m)} H_m V_i^{(m)}$$

and where $\{H_m\}$ is any sequence of matrices of $\Phi_{(k,k)}$ which converges to the zero matrix in terms of a metric topology associated with the metrization of $\Phi_{(k,k)}$ via the selection of one of the several norms that can be introduced in $\Phi_{(k,k)}$.

It appears that the Rinehart matrix derivative may be related to the Hausdorff matrix differential just as the Dwyer-Macphail matrix derivative is related to the Frechet-Gateaux matrix differential. One difference between the Rinehart matrix derivative and the Dwyer-Macphail matrix derivative, of course, is that the former is defined only for functions of square matrices while the latter matrix derivatives are defined for rectangular matrices as well.

We do not pursue this question concerning the relationships between various matrix derivatives or related investigations further. Rather we have mentioned these other notions of abstract differentials and matrix derivatives only to call the reader's attention to the fact that our discussion in no way exhausts all possible approaches to the study of matrix derivatives and matrix differentials.

3.2. BASIC ALGEBRAIC AND TOPOLOGICAL CONCEPTS

In this section we explore certain fundamental algebraic and topological concepts relevant to the notion of abstract spaces. Later we examine the idea of a differential of an abstract function or operator whose domain and range are real or complex Banach spaces or Banach algebras, and this study requires that we have a summary of some of the alge-

braic and topological properties of these two types of abstract spaces.

Consider a set S with at least two distinct elements which together with two operations from $S \times S$ to S called "addition" and "scalar multiplication" satisfy the following axioms:

A.1. S is an additive abelian group;

A.2. Scalar multiplication is such that (a), to every scalar λ in Φ (the real or complex field) and every x in S there corresponds a unique "scalar product" $\lambda \cdot x$ in S with $\lambda \cdot x = x \cdot \lambda$; (b), $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$ and $(\alpha\beta) \cdot x = \alpha \cdot (\beta \cdot x)$ for every α and β in Φ and every x in S ; (c), $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$ for every λ in Φ and every x and y in S ; and (d), $1 \cdot x = x$ for every x in S .

Then S is called a "linear (vector) space" (or a linear system or module) over the scalar field Φ of real or complex numbers and its elements are called "vectors." If in addition there is a third operation from $S \times S$ to S called "multiplication" which satisfies the additional axioms 3.e to 3.g below, then S is called a "linear associative algebra" over the scalar field Φ of real or complex numbers:

A.3. Multiplication of the elements of S leads to a unique "product" such that (e), $(xy)z = x(yz)$ for every x, y and z in S ; (f), $x(y+z) = xy+xz$ and $(x+y)z = xz+yz$ for every x, y and z in S ; and (g), $(\alpha x) \cdot (\beta y) = (\alpha\beta) \cdot (xy)$ for every α and β in Φ and x and y in S .

We say that S is an "algebra with a unit element" if there exists an element e such that $ex = x = xe$ for every x in S ; and we say that S is a "commutative" or "abelian" algebra if $xy = yx$ for every x and y in

S. Moreover, we simply say that "S is a real (or complex) vector space or algebra" rather than saying that "S is a vector space or algebra over the real (or complex) scalar field Φ ."

For our purposes we need not develop the general algebraic theory associated with either of these two types of algebraic systems. Instead we are primarily concerned here with certain parts of a theory of abstract functions in which the domains and ranges of the functions are such algebraic systems as a vector space or algebra in which a suitable topology also has been introduced. On the other hand, for this purpose, algebraic notions cannot be completely ignored since concepts of analysis originate from algebraic operations by passage to a limit.

A linear space or an algebra S is called a "normed linear space" or a "normed algebra" whenever there corresponds to every x in S a unique real number $\|x\|$ called the "norm of x" with these properties:

N.1. $\|x\| \geq 0$ for every x in S;

N.2. $\|x\| = 0$ if and only if $x = \theta$ where θ is the unit or zero of S as an additive group;

N.3. $\|\lambda \cdot x\| = |\lambda| \cdot \|x\|$ for every λ in Φ and x in S where $|\lambda|$ denotes the absolute value of the scalar λ in the usual sense; and

N.4. $\|x+y\| \leq \|x\| + \|y\|$ for every x and y in S.

Given that S is a normed vector space or normed algebra over Φ we may introduce a "metric" d which gives the "distance" $d(x,y)$ between elements of S by considering the function $d(x,y) = \|x-y\|$ from SXS to R. We find that this metric satisfies the following properties:

- M.1. $d(x,y) \geq 0$ for every x and y in S ;
- M.2. $d(x,y) = 0$ if and only if $x = y$;
- M.3. $d(x,y) = d(y,x)$ for every x and y in S ; and
- M.4. $d(x,y) \leq d(x,z) + d(z,y)$ for every x,y and z in S .

In a normed space S the concepts of "limit of a sequence" and of "Cauchy sequence" or "fundamental sequence" can be introduced in terms of the metric arising from the norm of the space; and S is called a "complete" space whenever each of its fundamental sequences converges to a limit in S . For convergent sequences of points of S the following well known theorems can be established:

Theorem 3.2.1. If a sequence of points of S converges to a point in S then every subsequence of this sequence converges to the same point;
and

Theorem 3.2.2. A sequence of points of S can converge to at most one limit; in other words, a convergent sequence has a unique limit.

In turn, in a normed space S , the notions of open and closed spheres with center x and radius r , of a neighborhood of a point x of S , of a bounded set, of a limit point of a set, of the closure of a set, of a closed set, of an open set, of a compact set, and other topological notions can be introduced in S in the usual manner. Thus if S is also a vector space or algebra, we arrive at a mathematical system which is simultaneously an algebraic system and a topological space.

When S is a normed linear vector space, from the relations

$$\|(x+y) - (x_n+y_n)\| \leq \|x-x_n\| + \|y-y_n\|$$

and

$$\|\lambda x - \lambda x_n\| = |\lambda| \|x - x_n\|$$

and

$$\|\lambda_n x - \lambda x\| = |\lambda_n - \lambda| \|x\|$$

and

$$|\|x\| - \|x_n\|| \leq \|x - x_n\|$$

we see that if $x_n \rightarrow x$, $y_n \rightarrow y$ and $\lambda_n \rightarrow \lambda$, then $x_n + y_n \rightarrow x + y$, $\lambda x_n \rightarrow \lambda x$, $\lambda_n x \rightarrow \lambda x$, $\lambda_n x_n \rightarrow \lambda x$ and $\|x_n\| \rightarrow \|x\|$. Thus the arithmetic operations of S as well as the norm of S are continuous functions of their arguments. And if S is a normed linear associative algebra over Φ and its norm satisfies the additional property,

N.5. $\|xy\| \leq \|x\| \cdot \|y\|$ for every x and y in S , from the relation

$$\|xy - x_n y_n\| \leq \|x\| \|y - y_n\| + \|y_n\| \|x - x_n\|$$

it immediately follows that whenever $x_n \rightarrow x$ or $y_n \rightarrow y$, then $x_n y \rightarrow xy$ or $xy_n \rightarrow xy$ or $x_n y_n \rightarrow xy$. Consequently, multiplication is a continuous operation too.

We are now in a position to define a real or complex Banach space and a real or complex Banach algebra:

Definition 3.2.1. A normed linear space S over the real (complex) field Φ which is complete in the metric topology based on its norm is called a real (complex) Banach space; and a normed linear associative

algebra S over Φ whose norm satisfies the property N.5 is called a real (complex) Banach algebra if it is complete as a metric space in its norm topology.

This completes our review of some of the fundamental algebraic and topological notions found in the theory of abstract spaces, particularly real or complex Banach spaces and Banach algebras. In Section 3.3 we proceed to introduce the Frechet differential and the Gateaux differential of an abstract function from and to Banach spaces or Banach algebras over Φ .

3.3. GATEAUX AND FRECHET DIFFERENTIALS OF ABSTRACT FUNCTIONS

It is not our purpose here to develop the theory of Gateaux and Frechet differentials for abstract functions either rigorously or in complete mathematical detail. Rather in the first part of this chapter, Sections 3.2, 3.3 and 3.4, we state some of the fundamental concepts of this theory in order to give a partial orientation for our main results which follow in Section 3.5. For all details and further information we refer the interested reader to those articles and books on this part of functional analysis that are listed in the Bibliography, namely Refs. [3, 36, 37, 38, 39, 40, 49, 50, 51, 59, 67, 75, 76, 102, 103, 104, 116, 134, 135, 146, 167, 168]. We do single out for special mention, however, the book by L. A. Lustenik and V. J. Sobolev [95] and the book by E. Hille and R. S. Phillips [60] because this summary has been influenced to a considerable degree by these two general and excellent works on functional analysis in abstract spaces.

Let X and Y be two real (or complex) Banach spaces or algebras and $y = f(x)$ an operator with domain D contained in X and range R contained in Y . We assume that D has a non-empty interior, and let x be an interior point of D and h an arbitrary non-zero element of X ; there exists a $\rho = \rho(x, h) > 0$ such that whenever λ is a scalar for which $|\lambda| \leq \rho$, then the element $x + \lambda h$ is contained in D too (e.g., take $\rho(x, h) < r_x \|h\|^{-1}$ where r_x is a radius of an open sphere about x that is properly contained in D).

Definition 3.3.1. The abstract function $y = f(x)$ on X to Y , real or complex Banach spaces or algebras, defined in a domain D of X having a non-empty interior, has a Gateaux differential $df(x; h)$ at an interior point x of D with increment h , provided for every positive ϵ there exists a positive δ such that

$$\|f(x + \lambda h) - f(x) - \lambda df(x; h)\| \leq \epsilon |\lambda|$$

whenever $|\lambda| \leq \delta$; or, equivalently, when

$$\lim_{\lambda \rightarrow 0} \left\| \frac{f(x + \lambda h) - f(x)}{\lambda} - df(x; h) \right\| = 0.$$

If $y = f(x)$ has a Gateaux differential at x for every increment h in X then $y = f(x)$ is said to be Gateaux differentiable at x ; and if $y = f(x)$ is Gateaux differentiable at every point of an open set D_* contained in D , it is said to be Gateaux differentiable on D_* .

To illustrate this notion of Gateaux differential let X be the ordinary euclidean plane and Y the real line. Then the real function

$y = f(x_1, x_2)$ of two real variables has a Gateaux differential at a point $x = (x_1, x_2)$ of an open set D_* in E^2 with increment $h = (h_1, h_2)$ if

$$df(x;h) = \lim_{t \rightarrow 0} \frac{f(x_1+th_1, x_2+th_2) - f(x_1, x_2)}{t}$$

exists. This obviously is equivalent to requiring that the directional derivative of $y = f(x_1, x_2)$ exists at $x = (x_1, x_2)$ for the direction vector $h = (h_1, h_2)$.

Clearly, if an abstract function possesses a Gateaux differential at x with increment h then $df(x;h)$ is unique. The Gateaux differential can be given another useful characterization as follows: For fixed but arbitrary elements x and h ,

$$u(\lambda) = f(x+\lambda h),$$

may be interpreted as an abstract function $z = u(\lambda)$ from Φ to Y ; and we may speak of the derivative $u'(\lambda) = \frac{d}{d\lambda} u(\lambda)$ of $z = u(\lambda)$ with respect to λ in the sense that

$$u'(\lambda) = \lim_{\Delta\lambda \rightarrow 0} \frac{u(\lambda+\Delta\lambda) - u(\lambda)}{\Delta\lambda}$$

is understood to exist as a limit in terms of norm convergence—in other words, if

$$\lim_{\Delta\lambda \rightarrow 0} \left\| \frac{u(\lambda+\Delta\lambda) - u(\lambda)}{\Delta\lambda} - u'(\lambda) \right\| = 0.$$

Then the Gateaux differential $df(x;h)$ exists and is continuous in x in the set $C_\rho(x, h) = \{x+\lambda h \mid |\lambda| \leq \rho\}$ if and only if $u(\lambda) = f(x+\lambda h)$ has a

continuous derivative $u'(\lambda)$ in the interval $[-\rho, \rho]$; and then,

$$df(x;h) = \lim_{\lambda \rightarrow 0} \frac{d}{d\lambda} f(x+\lambda h) = f'(x+\lambda h) \Big|_{\lambda=0} .$$

In addition, it can be shown that whenever the Gateaux differential of an abstract function $y = f(x)$ exists at x with increment h and is continuous in x , then $df(x;h)$ is a homogeneous and additive function in h ; i.e.,

$$df(x; \lambda_1 h_1 + \lambda_2 h_2) = \lambda_1 df(x; h_1) + \lambda_2 df(x; h_2) .$$

The notion of Frechet differential of an abstract function is not quite as general as the notion of Gateaux differential as the following definition indicates;

Definition 3.3.2. An abstract function $y = f(x)$ on X to Y , real or complex Banach spaces or algebras, defined in a domain D of X having a non-empty interior, is said to be Frechet differentiable at an interior point x of D provided there exists an abstract function $\delta f(x;h)$ of h from X to Y and depending also on x with the following properties: (a) $\delta f(x;h)$ is a continuous, linear function in h ; and (b) for every position ϵ there exists a positive η such that whenever $\|h\| \leq \eta$,

$$\|f(x+h) - f(x) - \delta f(x;h)\| \leq \epsilon \|h\| ;$$

or, equivalent, when

$$\lim_{\|h\| \rightarrow 0} \frac{1}{\|h\|} \|f(x+h) - f(x) - \delta f(x;h)\| = 0 .$$

When $\delta f(x;h)$ exists it is called a Frechet differential of $y = f(x)$ at x .

Again, as with the Gateaux differential, a Frechet differential is unique. Moreover, as the following theorem indicates, provided certain regularity conditions are satisfied, the Frechet and Gateaux differential are co-existent and co-equal.

Theorem 3.3.1. If the Frechet differential $\delta f(x;h)$ of an abstract function $y = f(x)$, exists, then a fortiori the Gateaux differential $df(x;h)$ exists and the two differentials are equal: $df(x,a) = \delta f(x;h)$. Conversely, if the Gateaux differential $df(x;h)$ exists and is continuous in x and h , then the Frechet differential $\delta f(x;h)$ exists and again the two differentials are equal: $\delta f(x;h) = df(x;h)$.

Proof. Suppose $\delta f(x;h)$ exists; then

$$f(x+h) - f(x) = \delta f(x;h) + \psi(x;h)$$

where $\|\psi(x;h)\| = o(\|h\|)$. Now since $\delta f(x;h)$ is homogeneous in h , we find that for an arbitrary scalar λ ,

$$\begin{aligned} f(x+\lambda h) - f(x) &= \delta f(x;\lambda h) + \psi(x;\lambda h) \\ &= \lambda \delta f(x;h) + \psi(x;\lambda h) . \end{aligned}$$

Consequently,

$$\left\| \frac{f(x+\lambda h) - f(x)}{\lambda} - \delta f(x;h) \right\| = \|h\| \left(\frac{\|\psi(x;\lambda h)\|}{\|\lambda h\|} \right) ;$$

and, therefore,

$$\lim_{\lambda \rightarrow 0} \left\| \frac{f(x+\lambda h) - f(x)}{\lambda} - \delta f(x; h) \right\| = \|h\| \lim_{\lambda \rightarrow 0} \left(\frac{\|\psi(x; \lambda h)\|}{\|\lambda h\|} \right) = 0.$$

Hence the Gateaux differential $df(x; h)$ exists and is simply $\delta f(x; h)$.

Next, suppose that the Gateaux differential $df(x; h)$ exists and is continuous in x and h . First we note that since $df(x; h)$ exists at x for arbitrary increments h , a fortiori for increments h_* having $\|h_*\| = 1$, to every positive ϵ and fixed x and h_* there corresponds a positive $\eta = \eta(\epsilon, x, h_*)$ such that

$$\left\| \frac{f(x + \lambda_* h_*) - f(x)}{\lambda_*} - df(x; h_*) \right\| \leq \epsilon$$

whenever $|\lambda_*| \leq \eta$. Now, for arbitrary h , write

$$f(x+h) = f(x + \lambda_* h_*)$$

where $\lambda_* = \|h\|$ and $h_* = h/\|h\|$ with $\|h_*\| = 1$. Then, since $df(x; h)$ is continuous in x , it is homogeneous in h , and

$$df(x; h) = df(x; \lambda_* h_*) = \lambda_* df(x; h_*);$$

and, therefore, for this particular choice of λ_* and h_* the inequality above leads to the conclusion that for every positive ϵ ,

$$\|f(x + \lambda_* h_*) - f(x) - df(x; \lambda_* h_*)\| \leq \epsilon |\lambda_*|,$$

whenever $|\lambda_*| \leq \eta_*$ and $\eta_* = \inf \{ \eta(\epsilon, x, h_*) \mid \|h_*\| = 1 \}$; and this is equivalent to

$$\lim_{\|h\| \rightarrow 0} \frac{1}{\|h\|} \|f(x+h) - f(x) - df(x;h)\| = 0.$$

Thus, $df(x;h)$ is a continuous, linear function in h (since it was assumed to be continuous in x and h) satisfying conditions (a) and (b) of Definition 3.3.2.; and since the Frechet differential is unique, $\delta f(x;h) = df(x;h)$.

We now summarize briefly some of the other important properties of the Frechet differential of an abstract function. Let $f(x)$, $u(x)$ and $v(x)$ be Frechet differentiable functions; then:

P.1. $f(x) = c$ identically in x (i.e., $f(x)$ is a constant function) if and only if $\delta f(x;h) = 0$ identically in h and x .

P.2. If $g(x) = u(x) + v(x)$, then $\delta g(x;h) = \delta u(x;h) + \delta v(x;h)$.

P.3. If $g(x) = u(x) * v(x)$, where $*$ is an associative multiplication, then $\delta g(x;h) = \delta u(x;h) * v(x) + u(x) * \delta v(x;h)$.

P.4. If $y = f(x)$ is non-singular at x (i.e., $y = f^{-1}(x)$ exists at x), then $f^{-1}(x;h) = f^{-1}(x) \delta f(x;h) f^{-1}(x)$.

P.5. If $g(y) = u(v(y))$, then $\delta g(y;h) = \delta u(x; \delta v(y;h))$ where $x = v(y)$.

The linear transformation $f'(x) = \delta f(x; \cdot)$ from X to Y has the property that there exists a positive scalar M such that

$$\|f'(x)h\| \leq M\|h\|$$

for all elements h of X since

$$\delta f(x;h) = f'(x)h$$

is continuous in h and therefore bounded in h ; in fact, any linear function from and to Banach spaces over Φ is continuous if and only if it is bounded. Thus $f'(x)$ is itself an element of the Banach space $L(X, Y)$ of all bounded linear transformations from X to Y where

$$\|T\| = \sup \{\|Tx\| \mid \|x\| \leq 1\}$$

is taken as the norm of an element $y = T(x)$ of $L(X, Y)$.

Defintion 3.2.3. The bounded linear transformation $f'(x)$ from X to Y is called the Frechet derivative at x of the Frechet differentiable function $y = f(x)$.

We can restate properties P.1 through P.5 in terms of $f'(x)$ rather than $\delta f(x; h)$ as follows:

- Q.1. $f(x) = c$ if and only if $f'(x) = 0$ identically in x ; and $\delta f(x; h) = 0$ identically in h if and only if $f'(x) = 0$.
- Q.2. If $g(x) = u(x) + v(x)$, then $g'(x) = u'(x) + v'(x)$.
- Q.3. If $g(x) = u(x) * v(x)$ where $*$ is an associative multiplication, then $g'(x) = u'(x) * v(x) + u(x) * v'(x)$.
- Q.4. If $y = f(x)$ is non-singular, then $(f^{-1}(x))' = -f^{-1}(x)f'(x)f^{-1}(x)$.
- Q.5. If $g(y) = u(v(y))$ then $g'(y) = u'(v(y))v'(y)$.

This concludes our review of some parts of the theory of first order Frechet and Gateaux differentials for abstract functions whose domain and ranges are real or complex Banach spaces or algebras. In the next section we proceed to give a short summary of some parts of the theory of

higher order Frechet and Gateaux differentials and this requires that we first introduce the notion of an "abstract polynomial."

3.4. ABSTRACT POLYNOMIALS AND HIGHER ORDER GATEAUX AND FRECHET DIFFERENTIALS

Before higher order Frechet differential can be introduced, we need a meaningful concept of "abstract polynomial." In elementary analysis, of course, a polynomial $P(x)$ is simply a sum of homogeneous forms in x of different degrees, while a homogeneous form in x of degree k is merely the product of a constant with x multiplied by itself k -times. But in a Banach space there is no ready made notion of "multiplication" upon which a theory of abstract polynomial like that found in the ordinary calculus can be easily constructed. Thus we must first search for an appropriate notion of "product" to serve as a basis for a theory of polynomials in real or complex Banach spaces.

Whenever $y = A(x)$ is a linear transformation from and to real or complex Banach spaces or Banach algebras, we may interpret $y = A(x)$ as a "product" $y = Ax$ formed by "multiplying" the elements A and x of two different normed linear spaces. Moreover, the product $y = Ax$ satisfies the following properties:

$$(A_1 + A_2)x = A_1x + A_2x; \quad A(x_1 + x_2) = Ax_1 + Ax_2;$$

$$(\alpha A)(\beta x) = (\alpha\beta)(Ax); \quad \text{and} \quad \|Ax\| \leq \|A\| \|x\|,$$

where we take $\|A\|$ to be the non-negative real number given as

$\sup\{\|Ax\| \mid \|x\| \leq 1\}$. In addition, when $A_n \rightarrow A$ and $x_n \rightarrow x$, A_n bounded,

then $A_n x \rightarrow Ax$, $Ax_n \rightarrow Ax$ and $A_n x_n \rightarrow Ax$; i.e., the product $y = Ax$ when restricted to bounded linear operators is a continuous function of its arguments A and x . This continuity property of the product $y = Ax$ follows as a consequence of Banach's theorem which states that a sequence of strongly convergent bounded operators has a bounded sequence of norms.

These properties justify our interpretation of $y = Ax$ as a "product" of the elements A and x . In fact, going one step further, we might even call $y = Ax$ a homogeneous form of degree one in x or an abstract homogeneous polynomial of degree one.

Accepting this possibility, why not interpret products of the form

$$y = Ax_1 x_2 \dots x_k$$

involving more than two factors as the image

$$y = A(x_1, x_2, \dots, x_k)$$

of a point $x = (x_1, x_2, \dots, x_k)$ under an abstract operator A which is a multilinear form in k -variables?

Next consider a symmetric k -linear form,

$$y = a^{(k)}(x_1, x_2, \dots, x_k) ;$$

$a^{(k)}$ is a multilinear form in k -variables in which each x_j for $1 \leq j \leq k$ is an element of the same abstract space and which has the same image

$$a^{(k)}(x_1, x_2, \dots, x_k) = a^{(k)}(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(k)})$$

for any point $x_\sigma = (x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(k)})$ obtained by permuting the components of $x = (x_1, x_2, \dots, x_k)$. We may obtain then a homogeneous form $a^{(k)}x^k$ of degree k in x , or a "product" of $a^{(k)}$ with x "multiplied" by itself k -times, by setting

$$a^{(k)}x^k = a^{(k)}(x, x, \dots, x)$$

where each of the k -variables of $a^{(k)}$ has been assigned a common value x :

$$x_1 = x_2 = \dots = x_k = x.$$

Finally, as in elementary analysis, a sum of abstract homogeneous forms in x of degree not exceeding k , say

$$P_k(x) = a^{(0)} + a^{(1)}x + \dots + a^{(k)}x^k,$$

is called an abstract polynomial of degree k in x whenever $a^{(k)}x^k \neq \theta$ identically in x .

It turns out that these notions presented here in summary form can be made to serve as a basis for establishing an interesting and fruitful mathematical theory of abstract polynomials in Banach spaces; but we do not pursue the development of this theory beyond the motivation presented here. Instead, once again, the interested reader is referred to the references listed in Section 3.3 and especially to [60] and [95] for all details.

Using this theory of abstract polynomial, we are now in a position

to introduce higher-order Frechet differentials for abstract functions. Thus let $y = f(x)$ be an operator from the real or complex Banach space or algebra X to another real or complex Banach space or algebra Y , and assume that $y = f(x)$ is defined on a domain D of X which has a non-empty interior. Then we say as in

Definition 3.4.1. The abstract function $y = f(x)$ is k-times Frechet differentiable at an interior point x of D provided there exists an abstract function $f_k(x;h)$ of h from X to Y and depending also on x such that (a) $f_k(x;h) = \sum_{j=1}^k a^{(j)} h^j$ is a continuous abstract polynomial of degree k in h ; and (b) for every positive ϵ there exists a positive η for which

$$\|f(x+h) - f(x) - f_k(x;h)\| \leq \epsilon \|h\|^k$$

whenever $\|h\| \leq \eta$; or, equivalently, such that

$$\lim_{\|h\| \rightarrow 0} \left(\frac{1}{\|h\|} \right)^k \|f(x+h) - f(x) - f_k(x;h)\| = 0.$$

When $f_k(x;h)$ exists it is called a k-th degree Taylor's sum of $y = f(x)$, about x and we denote $n! a^{(k)} h^k$ by $\delta^k f(x;h) = f^{(k)}(x) h^k$ and call this homogeneous polynomial of degree k in h a k-th order Frechet differential at x and $f^{(k)}(x)$ its k-th order Frechet derivative at x .

It can be shown that if $y = f(x)$ is k -times Frechet differentiable at x having a k -th order Frechet differential $\delta^k f(x;h) = f^{(k)}(x) h^k$ there, then for $1 \leq m \leq k$, $y = f(x)$ is also m -times Frechet differentiable having an m -th order Frechet differential $\delta^m f(x;h) = f^{(m)}(x) h^m$ at x .

Thus any k-th degree Taylor's sum of $y = f(x)$ can be written in the form

$$f_k(x;h) = \sum_{m=1}^k \frac{1}{m!} \delta^m f(x;h)$$

or, equivalently, in the form

$$f_k(x;h) = \sum_{m=1}^k \frac{1}{m!} f^{(m)}(x) h^m .$$

Moreover, a Taylor's sum and Frechet differential of a k-times Frechet differentiable abstract function are unique. In addition, whenever $\delta^k f(x;h)$ and $\delta^{k+1} f(x;h)$ both exist, then

$$\lim_{\|h\| \rightarrow 0} \frac{1}{\|h\|} \|\delta^k f(x+h;h) - \delta^k f(x;h) - \delta^{k+1} f(x;h)\| = 0 ;$$

in other words, the first order Frechet differential of a Frechet differential (interpreted as an abstract function of x) is merely the Frechet differential of next highest order of the given abstract function.

In introducing higher order Gateaux differentials of abstract functions, we find it convenient to denote the first order Gateaux differential of $y = f(x)$ by the notation $df(x;h)$ or $d_x^h f(x)$ and to use the two designations for first order Gateaux differentials interchangeably. Now for fixed but arbitrary $h = h_1$ the Gateaux differential $df(x;h_1)$ is called the "first variation at x " with increment h_1 of the abstract function $y = f(x)$. If for each h_1 , the abstract function $df(x;h_1)$ of x is Gateaux differentiable at x (and therefore $d_x^{h_1} df(x;h_1)$ exists for every h_1 and h), then for fixed but arbitrary $h = h_2$, $d_x^{h_2} df(x;h_1)$ is called

the "second variation at x " with increments h_1 and h_2 of the abstract function $y = f(x)$ and denoted by $d^2f(x;h_1,h_2)$. More generally, provided $d^{k-1}f(x;h_1,h_2,\dots,h_{k-1})$ exists for fixed but arbitrary increments and as an abstract function of x is Gateaux differentiable at x for all these increments (so that $d_x^h df(x;h_1,h_2,\dots,h_{k-1})$ exists for every $h_1,h_2,\dots,h_{k-1}$ and h), then for fixed but arbitrary $h = h_k$, $d_x^{h_k} d^{k-1}f(x;h_1,h_2,\dots,h_{k-1})$ is called the " k -th order variation at x " with increments $h_1,h_2,\dots$, and h_k of the abstract function $y = f(x)$ and denoted by $d^kf(x;h_1,h_2,\dots,h_k)$.

Definition 3.4.2. A k -th order variation $d^kf(x;h_1,h_2,\dots,h_k)$ of the abstract function $y = f(x)$ whose increments are assigned a common value $h_1 = h_2 = \dots = h_k = h$ is called a k -th order Gateaux differential at x of the abstract function $y = f(x)$ and denoted by $d^kf(x;h)$; in other words,

$$d^kf(x;h) = d^kf(x;h,h,\dots,h)$$

where the element h appears k -times in the expression for the k -th order variation on the right hand side.

Let $\rho = (\rho_1,\rho_2,\dots,\rho_k)$ and denote by $C_\rho(x;h_1,h_2,\dots,h_k)$ the collection of all elements of the form $x + \sum_{i=1}^k \lambda_i h_i$ where $|\lambda_i| \leq \rho_i$ for $1 \leq i \leq k$. Then the k -th order variation $d^kf(x;h_1,h_2,\dots,h_k)$ exists and is continuous in $C_\rho(x;h_1,h_2,\dots,h_k)$ if and only if

$$\frac{\partial^k}{\partial \lambda_1 \partial \lambda_2 \dots \partial \lambda_k} f \left(x + \sum_{i=1}^k \lambda_i h_i \right)$$

exists and is continuous for $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$ when $|\lambda_i| \leq \rho_i$ for $1 \leq i \leq k$. Also, when these mixed partial derivatives are independent of the order of differentiation,

$$d^k f(x; h_1, h_2, \dots, h_k) = \frac{\partial^k}{\partial \lambda_1 \partial \lambda_2 \dots \partial \lambda_k} f \left(x + \sum_{i=1}^k \lambda_i h_i \right) \Bigg|_{\lambda_1 = \lambda_2 = \dots = \lambda_k = 0};$$

and if $h_1 = h_2 = \dots = h_k = h$,

$$d^k f(x; h) = \frac{d^k}{d\lambda^k} f(x + \lambda h) \Bigg|_{\lambda=0}.$$

As in the case of first order Gateaux differentials, moreover, a k -th order Gateaux differential is unique and if $d^k f(x; h)$ exists and is continuous in x then $d^k f(x; h)$ is a homogeneous polynomial of degree k in h .

Having introduced higher order Frechet and Gateaux differential of abstract functions, we now state a key result insofar as the purposes of this summary are concerned which shows that under certain regularity conditions higher order Frechet and Gateaux differentials co-exists and are co-equal too.

Theorem 3.4.1. If an abstract function $y = f(x)$ has a k -th order Frechet differential $\delta^k f(x; h)$ and the remainder term $\psi(x; \lambda h)$ of its k -th degree Taylor's sum has a continuous derivative in a neighborhood of $\lambda = 0$ for fixed x and h , then $y = f(x)$ also has a k -th order Gateaux differential $d^k f(x; h)$ which is continuous in x and h and $d^k f(x; h) = \delta^k f(x; h)$. Conversely, if $y = f(x)$ has a k -th order Gateaux differential and $d^k f(x; h)$ is continuous in x and h , then $y = f(x)$ has a k -th order Frechet differ-

ential and $\delta^k f(x;h) = d^k f(x;h).$

Proof. A proof of this theorem would take us far beyond our immediate interests and the reader is referred to [95] where on pages 188 to 191 this theorem is rigorously established.

3.5. A BASIC RELATION BETWEEN DWYER-MACPHAIL MATRIX DERIVATIVES AND FRECHET-GATEAUX MATRIX DIFFERENTIALS

Returning to our main topic in this section of Chapter 3, we present new extensions of the theory of Dwyer-Macphail matrix derivatives in the form of connections between these matrix derivatives and Frechet-Gateaux matrix differentials. Before we can give a precise statement of these results, however, and furnish rigorous proofs, we first need to turn the collection $\Phi_{(u,v)}$ of uxv -matrices over Φ into a complete normed linear space. We may do this in many ways since the matrix space $\Phi_{(u,v)}$ can be furnished with several norms. The interested readers are referred to the work of J. Tysver in [148] for a discussion of different matrix norms that have been suggested for the treatment of matrix errors. Tysver also gives a fairly complete list of references on matrix norms current to 1958.

From a topological point of view these various norms for $\Phi_{(u,v)}$ are equivalent as the following theorem indicates: Each n -dimensional topological linear vector space over the real or complex numbers is isomorphic with n -dimensional real or complex Euclidean space. A proof of this theorem can be found in [60] on page 13. Thus for our purposes,

we need not be concerned with the selection of a particular norm in $\Phi_{(u,v)}$ and our choice of the norm

$$\|A\| = \max_{(s,t)} |a_{st}| \quad \text{for } A = [a_{st}] \text{ in } \Phi_{(u,v)}$$

presents no restrictions. The theorems below whose proofs are based on the use of this specific norm remain valid, although some details might be modified, should another norm be chosen for $\Phi_{(u,v)}$ since any two norms for this matrix space have metric topologies associated with them which are topologically equivalent.

We leave the verification that $\Phi_{(u,v)}$ together with the norm $\|A\|$ satisfy the axioms for a Banach space over Φ to the reader. When $u = v = w$, then $\Phi_{(w,w)}$ with the same norm $\|A\|$ is a Banach algebra over Φ provided we note that

$$\|AB\| \leq w\|A\| \|B\|$$

and that this involves a slight modification of N.5 (one of the axioms for a Banach algebra presented in Section 3.2) which is inconsequential within the context of finite dimensional vector space like $\Phi_{(w,w)}$.

We now begin our main discussion, not for the general case, but initially for first order Dwyer-Macphail matrix derivatives and first order Frechet-Gateaux matrix differentials.

Theorem 3.5.1. Let $Y = F(X)$ be a matrix function from $X \in \Phi_{(m,n)}$ to $Y \in \Phi_{(p,q)}$ each of whose pq-element function $f_{rs}(f) = \langle F(X) \rangle_{rs}$ possesses

continuous first order partial derivatives with respect to each of the
mn-components x_{ij} of x . If $Z = F_X(H)$ is a matrix function from $\text{He}\Phi_{(m,n)}$
to $\text{Xe}\Phi_{(p,q)}$ satisfying

$$\lim_{\|H\| \rightarrow 0} \frac{1}{\|H\|} \|F(X+H) - F(X) - F_X(H)\| = 0 ,$$

then

$$F_X(H) = \left[\text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} H^T \right) \right]_{(r,s)} .$$

In addition, $F_X(H)$ is additive and homogeneous of degree one in H and
continuous in X and H ; and, moreover, $Y = F(X)$ is Frechet-Gateaux dif-
ferentiable and its Frechet-Gateaux differential $\delta F(X;H) = F^{(')}(X)H$ is
given by the linear transformation

$$F^{(')}(X) = \delta F(X; \cdot) = F_X(\cdot)$$

which is an abstract homogeneous matrix polynomial of degree one uniquely
determined by the first order Dwyer-Macphail matrix derivatives $\frac{\partial \langle F \rangle_{rs}}{\partial X}$
of $Y = F(X)$.

Proof. Recalling that J_{ij} is the $m \times n$ -matrix with one in its (i,j) -th place and zero elsewhere and K_{rs} is the $p \times q$ -matrix with one in its (r,s) -th place and zero elsewhere, let us write

$$H = \sum_{i=1}^m \sum_{j=1}^n h_{ij} J_{ij} = [h_{ij}] ,$$

$$F(X) = \sum_{r=1}^p \sum_{s=1}^q F_{rs}(x) K_{rs} = [F_{rs}(x)] ;$$

$$F(X+H) = \sum_{r=1}^p \sum_{s=1}^q F_{rs}(x+h) K_{rs} = [F_{rs}(x+h)] ,$$

and

$$F_X(H) = \sum_{r=1}^p \sum_{s=1}^q F_{rs}^{(')}(x;h) K_{rs} = [F_{rs}^{(')}(x;h)] .$$

Now using these expressions together with the definition of $\|A\|$ as a maximum element of the matrix A , the condition

$$\lim_{\|H\| \rightarrow 0} \frac{1}{\|H\|} \|F(X+H) - F(X) - F_X(H)\| = 0$$

can be stated as

$$\lim_{\max_{(i,j)} |h_{ij}| \rightarrow 0} \frac{\max_{(r,s)} |F_{rs}(x+h) - F_{rs}(x) - F_{rs}^{(')}(x;h)|}{\max_{(i,j)} |h_{ij}|} = 0 .$$

Consequently, for arbitrary indices r and s we find that

$$\lim_{\max_{(i,j)} |h_{ij}| \rightarrow 0} \frac{|F_{rs}(x+h) - F_{rs}(x) - F_{rs}^{(')}(x;h)|}{\max_{(i,j)} |h_{ij}|} = 0 .$$

Thus from the ordinary calculus of a scalar function of several scalar variables we know that $F_{rs}^{(')}(x;h)$ is simply the total differential of the function $y_{rs} = F_{rs}(x)$ and, therefore,

$$F_{rs}^{(')}(x;h) = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial F_{rs}(x)}{\partial x_{ij}} h_{ij} .$$

Hence, according to Theorem 2.6.3.,

$$F_{rs}^{(')}(x;h) = \text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} H^T \right)$$

which proves that

$$F_X(H) = \left[\text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} H^T \right) \right]_{(r,s)} .$$

From this expression for $F_X(H)$ we see immediately that $F_X(H)$ is additive and homogeneous of degree one in H and continuous in X and H . Thus from Definition 3.3.2 and the uniqueness of the Frechet differential we conclude that $\delta F(X;H)$ exists and in fact that $\delta F(X;H) = F_X(H)$. Consequently, according to Theorem 3.3.1, the Gateaux differential $dF(X;H)$ of $F(X)$ exists and $dF(X;H) = F_X(H)$ too. Thus $Y = F(X)$ is Frechet-Gateaux differentiable and writing $\delta F(X;H) = F^{(')}(X)H$ it follows that

$$F^{(')}(X) = \left[\text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} (\cdot)^T \right) \right] = F_X(\cdot) .$$

This completes the proof of Theorem 3.5.1.

Since $F^{(')}(X) = F(X; \cdot)$ is a linear transformation between finite dimensional vector spaces we may speak of its "matrix" $[F^{(')}(X)]$ whenever bases are selected in $\Phi_{(m,n)}$ and $\Phi_{(p,q)}$ and their elements given an effective ordering. In particular, for the bases $\{J_{ij}\}$ in $\Phi_{(m,n)}$ and $\{K_{rs}\}$ in $\Phi_{(p,q)}$, ordering their elements lexicographically, we seek the relationship between $[F^{(')}(X)]$ and the Dwyer-Macphail matrix derivatives $\frac{\partial \langle F \rangle_{rs}}{\partial X}$ of $Y = F(X)$. In addition we inquire whether the interpretation

for linear transformations of $\delta F(X;H) = F^{(')}(X)H$ as the "product" of the elements $F^{(')}(X)$ and H can be translated into a meaningful matrix statement of the form $[F^{(')}(X)H] = [F^{(')}(X)][H]$ where $[F^{(')}(X)H]$, $[F^{(')}(X)]$ and $[H]$ denote, respectively, the "matrices" of $F^{(')}(X)H$, $F^{(')}(X)$ and H with respect to the basis $\{J_{ij}\}$ for $\Phi_{(m,n)}$ and the basis $\{K_{rs}\}$ for $\Phi_{(p,q)}$. Then we would preserve for differentiable matrix functions, in either the language of linear transformations or matrices, the same relationship that holds between differential and derivatives in the ordinary calculus; and we could say in terms of linear transformations or matrices that the Frechet-Gateaux differential of a matrix function is the product of a derivative (uniquely determined by the function's Dwyer-Macphail matrix derivative) and an arbitrary incremental element of a matrix Banach space or algebra.

This discussion poses a subtle conceptual distinction, however, involving the difference between a "matrix" of an element of a vector space of matrices and the matrix element itself; and we strive to keep this distinction between A as a matrix and (as an element of a vector space) its representation $[A]$ with respect to a given basis constantly in focus through the following formal device: We associate with each matrix A in $\Phi_{(u,v)}$ the column vector $a^\sigma = A_\sigma = \sigma(A)$ in Φ_{uv} (the collection of column vectors of order uv over Φ) formed by arranging the columns of A underneath one another in lexicographical order; in other words

$$A_\sigma^T = (\{A\}_1^T, \{A\}_2^T, \dots, \{A\}_v^T) = \sigma^T(A)$$

where $\{A\}_k$ denotes the k -th column of A which is a $ux1$ -column vector.

Now suppose $Y = TX$ is a linear transformation from $X \in \Phi_{(u,v)}$ to $Y \in \Phi_{(p,q)}$. We can define a linear transformation

$$y^\sigma = T_\sigma(x) \quad \text{from } x^\sigma \in \Phi_{mn} \quad \text{to } y^\sigma \in \Phi_{pq}$$

by means of T and σ as follows:

$$y^\sigma = Y_\sigma = (TX)_\sigma = T_\sigma X_\sigma = T_\sigma(x^\sigma) .$$

Corresponding to the bases $E = \{E_{ij}\}$ in $\Phi_{(m,n)}$ and $F = \{F_{rs}\}$ in $\Phi_{(p,q)}$ we obtain the basis $e^\sigma = \{e_{ij}^\sigma = \sigma(E_{ij})\}$ in Φ_{mn} and $f^\sigma = \{f_{rs}^\sigma = \sigma(F_{rs})\}$ in Φ_{pq} . Each element $T_\sigma(e_{ij}^\sigma)$ can be expressed uniquely as a linear combination of the elements of the basis f^σ , say

$$T_\sigma e_{ij}^\sigma = \sum_{r=1}^p \sum_{s=1}^q t_{rs}^{ij} f_{rs}^\sigma .$$

Following a lexicographical ordering of the bases elements, we form the matrix $[T_\sigma]$ of T_σ with respect to the bases e^σ and f^σ by writing the coefficients appearing in this linear combination as the (i,j) -th column of $[T_\sigma]$. We note that many authors use these coefficients to form the (i,j) -th row of the matrix of a linear transformation, but we find it convenient to form columns instead of rows. Then, since

$$X_\sigma = \sum_{i=1}^m \sum_{j=1}^n x_{ij} e_{ij}^\sigma ,$$

we find that

$$\begin{aligned}
T_{\sigma} X_{\sigma} &= T_{\sigma} \left(\sum_{i=1}^m \sum_{j=1}^n x_{ij} e_{ij}^{\sigma} \right) \\
&= \sum_{i=1}^m \sum_{j=1}^n x_{ij} T_{\sigma}(e_{ij}^{\sigma}) \\
&= \sum_{r=1}^p \sum_{s=1}^q \left(\sum_{i=1}^m \sum_{j=1}^n x_{ij} t_{rs}^{ij} \right) f_{rs}^{\sigma} ;
\end{aligned}$$

and, therefore, whenever

$$Y_{\sigma} = T_{\sigma} X_{\sigma}$$

we see that

$$[Y_{\sigma}] = [T_{\sigma}] \cdot [X_{\sigma}]$$

where this matrix product is formed according to the standard rules for matrix multiplication.

Returning to the linear transformation $Y = TX$ between the matrix spaces $X \in \Phi_{(m,n)}$ and $Y \in \Phi_{(p,q)}$ we define the "matrix" $[X]$ of X with respect to the basis $E = \{E_{ij}\}$ for $\Phi_{(m,n)}$ as the matrix $[X_{\sigma}]$, the "matrix" $[Y]$ of Y or $[TX]$ of TX with respect to the basis $F = \{F_{rs}\}$ for $\Phi_{(p,q)}$ as the matrix $[Y_{\sigma}]$ or $[T_{\sigma}X_{\sigma}]$, and the matrix $[T]$ of T with respect to the bases E and F as the matrix $[T_{\sigma}]$. Consequently, in terms of this notion of a "matrix" of an element of a matrix space and this notion of a "matrix" of a linear transformation between matrix spaces, we find that the expression $Y = TX$ involving linear transformation can be translated into an equivalent expression of the same form involving matrices, namely

$$[Y] = [T][X] = [TX]$$

whenever bases are selected in the domain and range spaces of T .

Observing these special interpretations, we now state

Theorem 3.5.2. If $F^{(1)}(X)H = \delta F(X;H)$ is the Frechet-Gateaux matrix differential of the matrix function $Y = F(X)$ from $X \in \Phi_{(m,n)}$ to $Y \in \Phi_{(p,q)}$ whose element functions possess continuous first order partial derivatives, then the matrix of the linear transformation $F^{(1)}(X) = \delta F(X; \cdot)$ with respect to the basis $\{J_{ij}\}$ for $\Phi_{(m,n)}$ and the basis $\{K_{rs}\}$ for $\Phi_{(p,q)}$ is given by

$$[F^{(1)}(X)] = \begin{bmatrix} \left(\frac{\partial \langle F \rangle_{11}}{\partial X}\right)_1 & \left(\frac{\partial \langle F \rangle_{11}}{\partial X}\right)_2 & \dots & \left(\frac{\partial \langle F \rangle_{11}}{\partial X}\right)_m \\ \left(\frac{\partial \langle F \rangle_{12}}{\partial X}\right)_1 & \left(\frac{\partial \langle F \rangle_{12}}{\partial X}\right)_2 & \dots & \left(\frac{\partial \langle F \rangle_{12}}{\partial X}\right)_m \\ \vdots & \vdots & & \vdots \\ \left(\frac{\partial \langle F \rangle_{pq}}{\partial X}\right)_1 & \left(\frac{\partial \langle F \rangle_{pq}}{\partial X}\right)_2 & & \left(\frac{\partial \langle F \rangle_{pq}}{\partial X}\right)_m \end{bmatrix}$$

and the matrix of $\delta F(X;H) = F^{(1)}(X)H$ is given by

$$[F^{(1)}(X)H] = [F^{(1)}(X)][H] .$$

In the expression for $[F^{(1)}(X)]$ the notation $\left(\frac{\partial \langle F \rangle_{rs}}{\partial X}\right)_k$ denotes the k -th row of the Dwyer-Macphail matrix derivative $\frac{\partial \langle F \rangle_{rs}}{\partial X}$ of $Y = F(X)$.

Proof. Identifying $F^{(')}(X)$ with the linear transformation T in our previous discussion we see immediately that

$$[F_O^{(')}(X)H_O] = [F_O^{(')}(X)][H_O]$$

and, therefore, the second part of our theorem is certainly valid.

Now from Theorem 3.5.1 we know that

$$F^{(')}(X)H = \left[\text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} H^T \right) \right]_{(r,s)}$$

and, hence,

$$[F_O^{(')}(X)][H_O] = \left[\text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} H^T \right) \right]_{\sigma}.$$

Multiplying the matrices $[F_O^{(')}(X)]$ and $[H_O]$ and equating corresponding elements in this matrix equality leads to the observation that

$$\sum_{i=1}^m \sum_{j=1}^n f_{rs}^{ij} h_{ij} = \text{Tr} \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} H^T \right)$$

holds identically in $H = [h_{ij}]$ where f_{rs}^{ij} is the element of the matrix $[F_O^{(')}(X)]$ in the $((i,j), (r,s))$ -th position. Taking $H = J_{ij}$, gives the equality

$$f_{rs}^{ij} = \left\langle \frac{\partial \langle F \rangle_{rs}}{\partial X} \right\rangle_{ij};$$

and, consequently, the (r,s) -th row of the matrix $[F_O^{(')}(X)]$ is merely

$$\left(\left(\frac{\partial \langle F \rangle_{rs}}{\partial X} \right)_1, \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} \right)_2, \dots, \left(\frac{\partial \langle F \rangle_{rs}}{\partial X} \right)_m \right).$$

This completes our proof of Theorem 3.5.2.

We now turn our attention to extending the basic relationship that exists between first order Dwyer-Macphail matrix derivatives and Frechet-Gateaux matrix differentials to higher order matrix derivatives and matrix differentials. As in the case of first order matrix derivatives and differentials, we find again that higher order Frechet-Gateaux matrix differentials are abstract homogeneous matrix polynomials, and that these abstract matrix polynomials are uniquely determined in terms of higher order Dwyer-Macphail matrix derivatives.

Theorem 3.5.3. Let $Y = F(X)$ be a matrix function from $X \in \Phi_{(p,q)}$ to $Y \in \Phi_{(p,q)}$ each of whose pq-element functions $f_{rs}(x) = \langle F(X) \rangle_{rs}$ possesses continuous pure or mixed k-th order partial derivatives with respect to the mn-components x_{ij} of x . If $Z = P_X^{(k)}(H)$ is a matrix function from $H \in \Phi_{(m,n)}$ to $Z \in \Phi_{(p,q)}$ satisfying

$$\lim_{\|H\| \rightarrow 0} \left(\frac{1}{\|H\|} \right)^k \|F(X+H) - F(X) - P_X^{(k)}(H)\| = 0 ,$$

then

$$P_X^{(k)}(H) = \sum_{m=1}^k \frac{1}{m!} F_X^{(m)}(X)$$

where

$$F_X^{(m)}(H) = \left[\text{Tr} \left(\frac{\partial^m \langle F \rangle_{rs}}{\partial X^m} (H^m)^T \right) \right]_{(r,s)} .$$

$P_X^{(k)}(H)$ and $F_X^{(m)}(H)$ for $1 \leq m \leq k$ are continuous in X and H ; and

$F_X^{(m)}(H_1, H_2, \dots, H_m)$ is additive and homogeneous of degree one in each of

its m-matrix variables separately. Moreover, $Y = F(X)$ is k-times Frechet-Gateaux differentiable and its k-th order Frechet-Gateaux matrix differential $\delta^k F(X; H) = F^{(k)}(X) H^k$ is given by the transformation

$$F^{(k)}(X) = \delta^k F(X; \cdot) = F_X^{(k)}(\cdot)$$

which is a homogeneous matrix polynomial of degree k uniquely determined by the k-th order Dwyer-Macphail matrix derivatives $\frac{\partial^k \langle F \rangle_{rs}}{\partial X^k}$ of $Y = F(X)$.

Proof. We prove this theorem by mathematical induction where the n-th induction hypothesis H_n for $1 \leq n \leq k$ is taken to be the statement of Theorem 3.5.3 for the positive integer $k = n$. Clearly for $n = 1$ the induction hypothesis H_1 is true since we have established H_1 as our Theorem 3.5.1. Now assume that H_n is true whenever $1 \leq n \leq k-1$; we show below that this implies that H_{n+1} also is valid. In particular, therefore, when $n = k-1$ it follows that H_k is true and thus Theorem 3.5.3 is established as a consequence of the following modification of the principle of mathematical induction: Suppose Q is a subset of the positive integers such that (a), 1 is in Q ; and (b), for a given positive integer k , whenever $n < k$ and n is in Q ; then $n+1$ is in Q too. Then Q contains all positive integers less than or equal to k . The interested reader may refer to [149] where an excellent discussion of the Mathematical Induction Principle can be found, including a proof of this finite version which we use here.

It remains for us to show that H_n implies H_{n+1} whenever $1 \leq n \leq k-1$.

Let

$$\Delta_X(H) = (n+1)! (P_X^{(n+1)}(H) - P_X^{(n)}(H))$$

and consider the identity

$$\begin{aligned} F_{X+H}^{(n)}(H) - F_X^{(n)}(H) - \Delta_X(H) &= F_{X+H}^{(n)}(H) - F_X^{(n)}(H) \\ &+ (n+1)! (\Delta_H F(X) - P_X^{(n)}(H)) - (n+1)! (\Delta_H F(X) - P_X^{(n+1)}(H)) \end{aligned}$$

where $\Delta_H F(X) = F(X+H) - F(X)$. Now taking $\epsilon_* = \epsilon/n+1$ we know that there exists an $\eta > 0$ such that whenever $\|H\| \leq \eta$,

$$\begin{aligned} \|F_{X+H}^{(n)}(H) - F_X^{(n)}(H) - \Delta_X(H)\| &\leq \|F_{X+H}^{(n)}(H) - F_X^{(n)}(H)\| \\ &+ \epsilon \|H\|^n (1 + \|H\|) ; \end{aligned}$$

or, equivalently,

$$\begin{aligned} \frac{1}{\|H\|} \|F_{X+H}^{(n)}(H) - F_X^{(n)}(H) - \Delta_X(H)\| &\leq \frac{1}{\|H\|} \|F_{X+H}^{(n)}(H) - F_X^{(n)}(H)\| \\ &+ \epsilon \|H\|^{n-1} (1 + \|H\|) . \end{aligned}$$

Now according to our induction hypothesis

$$F_X^{(n)}(H) = \left[\text{Tr} \left(\frac{\partial^n \langle F(X) \rangle_{rs}}{\partial X^n} (H^n)^T \right) \right]_{(r,s)} ;$$

and, therefore,

$$\|F_{X+H}^{(n)}(H) - F_X^{(n)}(H)\|$$

$$= \max_{(r,s)} \left| \text{Tr} \left(\frac{\partial^n \langle F(X+H) \rangle_{rs}}{\partial X^n} - \frac{\partial^n \langle F(X) \rangle_{rs}}{\partial X^n} \right) (H^n)^T \right|.$$

If this maximum is assumed at (r_0, s_0) , then

$$\|F_{X+H}^{(n)}(H) - F_X^{(n)}(H)\|$$

$$= \left| \text{Tr} \left(\frac{\partial^n \langle F(X+H) \rangle_{r_0 s_0}}{\partial X^n} - \frac{\partial^n \langle F(X) \rangle_{r_0 s_0}}{\partial X^n} \right) (H^n)^T \right|$$

$$\leq mn \max_{(i,j)} \left| \langle \frac{\partial^n \langle F(X+H) \rangle_{r_0 s_0}}{\partial X^n} - \frac{\partial^n \langle F(X) \rangle_{r_0 s_0}}{\partial X^n} \rangle_{ij} \right| \|H\|^n.$$

Consequently,

$$\|F_{X+H}^{(n)}(H) - F_X^{(n)}(H)\| \leq K(H) \|H\|^n$$

where

$$K(H) = mn \max_{(i,j)} \left| \langle \frac{\partial^n \langle F(X+H) \rangle_{r_0 s_0}}{\partial X^n} - \frac{\partial^n \langle F(X) \rangle_{r_0 s_0}}{\partial X^n} \rangle_{ij} \right|$$

and $K(H) \rightarrow 0$ as $\|H\| \rightarrow 0$ since $\langle \frac{\partial^n \langle F(X) \rangle_{r_0 s_0}}{\partial X^n} \rangle_{ij}$ is a continuous function of X . Hence we obtain the inequality

$$\frac{1}{\|H\|} \|F_{X+H}^{(n)}(H) - F_X^{(n)}(H) - \Delta_X(H)\| \leq K_*(H) \|H\|^{n-1}$$

for sufficiently small $\|H\|$, where

$$K_*(H) = K(H) + \epsilon(\|H\|+1) \quad \text{and} \quad K_*(H) \rightarrow \epsilon \quad \text{as} \quad \|H\| \rightarrow 0.$$

From this it follows that whenever $1 \leq n \leq k-1$,

$$\lim_{\|H\| \rightarrow 0} \frac{1}{\|H\|} \|F_{X+H}^{(n)}(H) - F_X^{(n)}(H) - \Delta_X(H)\| = 0 ;$$

and, hence, according to Theorem 3.5.1,

$$\Delta_X(H) = \left[\text{Tr} \left(\frac{\partial < F_X^{(n)}(H) >_{rs}}{\partial X} H^T \right) \right]_{(r,s)} .$$

But from our induction hypothesis we know that

$$< F_X^{(n)}(H) >_{rs} = \text{Tr} \left(\frac{\partial^n < F >_{rs} (H^n)^T}{\partial X^n} \right) ;$$

and, according to Theorem 2.6.3,

$$\text{Tr} \left(\frac{\partial^n < F >_{rs} (H^n)^T}{\partial X^n} \right) = \sum \frac{\partial^n f(x)}{\partial x_{\sigma_1}^{k_1} \partial x_{\sigma_2}^{k_2} \dots \partial x_{\sigma_{mn}}^{k_{mn}}} h_{\sigma_1}^{k_1} h_{\sigma_2}^{k_2} \dots h_{\sigma_{mn}}^{k_{mn}}$$

where $\sum$ is taken over all $k = (k_1, k_2, \dots, k_{mn})$ such that $k_j \geq 0$ and $\sum_{j=1}^{mn} k_j = n$. Thus we find that

$$\begin{aligned} & \text{Tr} \left(\frac{\partial < F_X^{(n)}(H) >_{rs}}{\partial X} H^T \right) \\ &= \sum h_{\sigma_1}^{k_1} h_{\sigma_2}^{k_2} \dots h_{\sigma_{mn}}^{k_{mn}} \text{Tr} \left(\frac{\partial}{\partial X} \left(\frac{\partial^n f(x)}{\partial x_{\sigma_1}^{k_1} \partial x_{\sigma_2}^{k_2} \dots \partial x_{\sigma_{mn}}^{k_{mn}}} \right) H^T \right) \\ &= \sum h_{\sigma_1}^{k_1} h_{\sigma_2}^{k_2} \dots h_{\sigma_{mn}}^{k_{mn}} \sum_{i=1}^{mn} \frac{\partial^{n+1} f(x)}{\partial x_{\sigma_1}^{k_1} \dots \partial x_{\sigma_i}^{k_i+1} \dots \partial x_{\sigma_{mn}}^{k_{mn}}} h_{\sigma_i} \\ &= \sum \frac{\partial^{n+1} f(x)}{\partial x_{\sigma_1}^{k_1^*} \partial x_{\sigma_2}^{k_2^*} \dots \partial x_{\sigma_{mn}}^{k_{mn}^*}} h_{\sigma_1}^{k_1^*} h_{\sigma_2}^{k_2^*} \dots h_{\sigma_{mn}}^{k_{mn}^*} \end{aligned}$$

where $k^* = (k_1^*, k_2^*, \dots, k_{mn}^*)$ satisfies $k_j^* \geq 0$ and $\sum_{j=1}^{mn} k_j^* = n+1$; i.e., k^* is an mn -termed partition of the integer $n+1$. Thus it follows from Theorem 2.6.3 that

$$\text{Tr} \left(\frac{\partial < F_X^{(n)}(H) >_{rs}}{\partial X} H^T \right) = \text{Tr} \left(\frac{\partial^{n+1} < F >_{rs}}{\partial X^{n+1}} (H^{n+1})^T \right) ;$$

and, therefore, we have established that

$$\Delta_X(H) = \left[\text{Tr} \left(\frac{\partial^{n+1} < F >_{rs}}{\partial X^{n+1}} (H^{n+1})^T \right) \right]_{(r,s)} .$$

Moreover, according to our induction hypothesis,

$$P_X^{(n)}(H) = \sum_{m=1}^n \frac{1}{m!} F_X^{(m)}(H) ,$$

where, for $1 \leq m \leq n$,

$$F_X^{(m)}(H) = \left[\text{Tr} \frac{\partial^m < F >_{rs}}{\partial X^m} (H^m)^T \right] ;$$

and, therefore, since

$$P_X^{(n+1)}(H) = P_X^{(n)}(H) + \frac{1}{(n+1)!} \Delta_X(H) ,$$

we find that

$$P_X^{(n+1)}(H) = \sum_{m=1}^{n+1} \frac{1}{m!} F_X^{(m)}(H)$$

where we have put $F_X^{(n+1)}(H) = \Delta_X(H)$. Thus we have established the desired representation for $P_X^{(n+1)}(H)$ assuming that it was given for $P_X^{(n)}(H)$.

From this result we see at once that $F_X^{(n+1)}(H)$ is continuous in X and H ; and since $P_X^{(n)}(H)$ is continuous in X and H , it follows that $P_X^{(n+1)}(H)$ as the sum of $P_X^{(n)}(H)$ and $\frac{1}{(n+1)!} F_X^{(n+1)}(H)$ is continuous in X and H too.

Next we see that $\Delta_X(H_1, H_2, \dots, H_{n+1})$ is additive and homogeneous of degree one in each of the $(n+1)$ -matrix variables separately. We have shown that $\Delta_X(H_1, H_2, \dots, H_{n+1})$ can be written in the form

$$\Delta_X(H_1, H_2, \dots, H_{n+1}) = \left[\text{Tr} \left(\frac{\partial \langle F_X^{(n)}(H_1, H_2, \dots, H_n) \rangle_{rs}}{\partial X} H_{n+1}^T \right) \right] .$$

From this expression we see that $\Delta_X(H_1, H_2, \dots, H_{n+1})$ is additive and homogeneous of degree one in H_{n+1} since, in general,

$$\text{Tr}(\alpha A)(\beta B) = \alpha \beta \text{Tr} AB$$

and

$$\text{Tr} A(B+C) = \text{Tr} AB + \text{Tr} AC \quad \text{and} \quad \text{Tr}(A+B)C = \text{Tr} AC + \text{Tr} BC .$$

Now from our induction hypothesis, we know that $F_X^{(n)}(H_1, H_2, \dots, H_n)$ is additive and homogeneous of degree one in each of its n -matrix variables separately; and, therefore,

$$\frac{\partial \langle F_X^{(n)}(H_1, H_2, \dots, H_n) \rangle_{rs}}{\partial X}$$

is additive and homogeneous of degree one in each of these n -matrix variables too. Consequently, by the same two properties of the Tr -operator stated above, it follows that

$$F_X^{(n+1)}(H_1, H_2, \dots, H_{n+1}) = \Delta_X(H_1, H_2, \dots, H_{n+1})$$

is additive and homogeneous of degree one in each of its $(n+1)$ -matrix variables separately.

Finally, our proof has established that $\delta^{n+1}F(X;H)$ exists and is the first order Frechet differential of $\delta^n F(X;H)$ considered as a function of X ; i.e.,

$$\delta^{n+1}F(X;H) = \delta_X^H \delta^n F(X;H) \quad .$$

But since $\delta_X^H \delta^n F(X;H)$ exists, according to Theorem 3.3.1, the first order Gateaux differential of $\delta^n F(X;H)$ also exists, and

$$d_X^H \delta^n F(X;H) = \delta_X^H \delta^n F(X;H) \quad .$$

Now, according to our induction hypothesis, $Y = F(X)$ is n -times Frechet-Gateaux differentiable and therefore $d^n F(X;H)$ exists and

$$d^n F(X;H) = \delta^n F(X;H) \quad .$$

Thus we see that

$$\begin{aligned} d^{n+1}F(X;H) &= d_X^H d^n F(X;H) \\ &= d_X^H \delta^n F(X;H) \\ &= \delta_X^H \delta^n F(X;H) \\ &= \delta^{n+1}F(X;H) \quad . \end{aligned}$$

Consequently, $Y = F(X)$ is $(n+1)$ -times Frechet-Gateaux differentiable and

the proof of Theorem 3.5.3 is completed.

In Chapters 4 and 5, we find ourselves concerned with problems involving the optimization of scalar functions of matrix variables, and so we present here several theorems featuring matrix derivatives and matrix differentials which involve necessary conditions for critical points or sufficient conditions for a maximum or a minimum of such scalar functions of a matrix variable. These theorems are obvious results and need not involve us in considerable detail, but we do present them here for the sake of completeness. Moreover, since they feature the functions first or second order Dwyer-Macphail matrix derivatives or its first or second order Frechet-Gateaux matrix differentials, these theorems provide us with useful illustrations of some of the theorems developed in this section. First we present a theorem involving necessary conditions for an unrestricted critical point.

Theorem 3.5.4. Let $y = f(X)$ be a functional from $X \in \Phi_{(m,n)}$ to $y \in \Phi$ whose representation $z = f(x)$ as a function of several scalar variables possesses continuous partial derivatives with respect to each of the mn -variables x_{ij} of X . If $y = f(X)$ has a critical point at the matrix point $X = X_0$, then its first order Dwyer-Macphail matrix derivative $\frac{\partial f}{\partial X}$ vanishes at $X = X_0$; i.e.,

$$\frac{\partial f}{\partial X_0} = \left. \frac{\partial f}{\partial X} \right|_{X=X_0} = 0$$

where 0 denotes the $m \times n$ -matrix element of $\Phi_{(m,n)}$ whose elements are all zero. Equivalently, the functions first order Frechet-Gateaux matrix

differential $\delta f(X;H)$ vanishes identically in H at $X = X_0$; i.e.,

$$\delta f(X_0;H) = \left. \delta f(X;H) \right]_{X=X_0} = 0$$

for arbitrary matrix increments H in $\Phi_{(m,n)}$.

Proof. Since $y = f(X)$ has a critical point at $X = X_0$, $z = f(x)$ has a critical point at $x = x_0$. But we know from the elementary calculus that

$$\left. \frac{\partial f}{\partial x_{ij}} \right]_{x_{ij}=x_{ij}^0} = 0$$

since $z = f(x)$ possesses continuous first order partial derivatives; and, consequently,

$$\left. \frac{\partial f}{\partial X} \right]_{X=X_0} = \left[\left. \frac{\partial f}{\partial x_{ij}} \right]_{x_{ij}=x_{ij}^0} \right] = 0 .$$

Now suppose the Frechet-Gateaux derivative $f^{(')}(X) = \delta f(X; \cdot)$ of $y = f(X)$ vanishes at $X = X_0$; in other words, $f^{(')}(X_0) = 0$ (i.e., $f(X_0)$ is that linear transformation for which $f^{(')}(X_0)H = \delta f(X_0;H) = 0$ identically in H). According to Theorem 3.5.2,

$$[f^{(')}(X_0)] = \left(\left(\frac{\partial f}{\partial X_0} \right)_1, \left(\frac{\partial f}{\partial X_0} \right)_2, \dots, \left(\frac{\partial f}{\partial X_0} \right)_m \right) .$$

Hence, $[f^{(')}(X_0)] = 0$ if and only if $\frac{\partial f}{\partial X_0} = 0$; or, equivalently,

$f^{(')}(X_0) = 0$ if and only if $\frac{\partial f}{\partial X_0} = 0$; or, equivalently, $\delta f(X_0;H) = 0$

identically in H if and only if $\frac{\partial f}{\partial X_0} = 0$. Thus the proof of Theorem

3.5.4 is completed.

We now state a second theorem featuring second order Dwyer-Macphail matrix derivatives or second order Frechet-Gateaux matrix differentials which gives sufficient conditions for a local maximum or a local minimum of a scalar function of a matrix variable:

Theorem 3.5.5. Let $y = f(X)$ be a functional from $X \in \Phi_{(m,n)}$ to $y \in \Phi$ whose representation $z = f(x)$ as a function of several scalar variables possesses continuous second order partial derivatives with respect to the mn -variables x_{ij} of x . If at $X = X_0$, the Dwyer-Macphail matrix derivative vanishes, then (a), $y = f(X)$ has a minimum at $X = X_0$ provided

$$\text{Tr} \left(\left[\frac{\partial^2 f}{\partial X^2} \right]_{X=X_0} (H^2)^T \right) > 0 ;$$

and (b), $y = f(X)$ has a maximum at $X = X_0$ provided

$$\text{Tr} \left(\left[\frac{\partial^2 f}{\partial X^2} \right]_{X=X_0} (H^2)^T \right) < 0 .$$

Each of these two statements is understood to hold identically for non-zero H in $\Phi_{(m,n)}$. Equivalently, $y = f(X)$ has a minimum at $X = X_0$ provided $\delta^2 f(X_0; H) > 0$ for every non-zero matrix H ; or $y = f(X)$ has a maximum at $X = X_0$ provided $\delta^2 f(X_0; H) < 0$ for every non-zero matrix H .

Proof. From the ordinary calculus we know that $z = f(x)$ has a minimum at $x = x_0$, when $f'(x_0) = 0$, provided

$$\sum_{(i,j)} \sum_{(k,l)} h_{ij} \left[\frac{\partial^2 f(x)}{\partial x_{ij} \partial x_{kl}} \right]_{(x_{ij}=x_{ij}^0, x_{kl}=x_{kl}^0)} h_{kl} > 0 ;$$

while $z = f(x)$ has a maximum at $x = x_0$, when $f'(x_0) = 0$, provided

$$\sum_{(i,j)} \sum_{(k,l)} h_{ij} \left[\frac{\partial^2 f(x)}{\partial x_{ij} \partial x_{kl}} \right]_{(x_{ij}=x_{ij}^0, x_{kl}=x_{kl}^0)} h_{kl} < 0.$$

These conditions are understood to hold for arbitrary scalars h_{ij} unless all the h_{ij} are zero. But we know from Theorem 2.6.3 that

$$\text{Tr} \left(\left[\frac{\partial^2 f}{\partial X^2} \right]_{X=X^0} (H^2)^T \right) = \sum_{(i,j)} \sum_{(k,l)} h_{ij} \left[\frac{\partial^2 f(x)}{\partial x_{ij} \partial x_{kl}} \right]_{(x_{ij}=x_{ij}^0, x_{kl}=x_{kl}^0)} h_{kl}$$

and thus the first part of this theorem has been established.

Finally, according to Theorem 3.5.3, when $y = f(X)$ is a scalar function of a matrix variable

$$\delta^2 f(X; H) = \text{Tr} \left(\left[\frac{\partial^2 f}{\partial X^2} \right] (H^2)^T \right),$$

and this fact establishes the second part of Theorem 3.5.5.

In the next two chapters, we undertake applications of the theory and techniques of Dwyer-Macphail matrix derivatives and Frechet-Gateaux matrix differentials presented in Chapters 2 and 3. These applications deal with certain estimation problems in mathematical statistics, and they lead to questions involving the maximization or minimization of quite complicated functionals from vector or matrix spaces which can be expressed as composite functions of other vector or matrix functions.

CHAPTER 4

APPLICATIONS TO ESTIMATION PROBLEMS IN STATISTICS

4.1. INTRODUCTION

We now apply the theory and techniques of matrix derivatives and matrix differentials as developed in Chapters 2 and 3 to the problem of estimating the unknown parameters of a stochastic structure representing a linear structural relationship having errors-in-variables and/or errors-in-equations. Before abstracting the essential features of this problem into a general statistical model, let us first familiarize ourselves with its details through an illustration showing that it is not only one of theoretical interest, but has significant import in the applications of statistics as well. The illustration which we discuss below is drawn from physics; and it is concerned with estimating the parameters of the temporal distribution law of electromagnetic radiation emitted from a falling body. Different applications involving this statistical model can be found in the other physical sciences and in the behavioral sciences too. In the area of econometrics, for example, the reader can find interesting applications of the error-in-variable and/or error-in-equation statistical model in papers by M Girshick and T. Haavelmo [48] and D. Suits [133]. Examples of the application of this statistical model in other physical sciences are presented by F. Acton in [1].

Proceeding with our example, it is known that the distribution law of infrared radiation emitted from a "black body" or complete radiator

can be found by integrating the body's spectral radiation distribution law over the interval (λ_1, λ_2) denoting the infrared wavelength spectrum; thus,

$$J = \int_{\lambda_1}^{\lambda_2} E d\lambda ,$$

where $E d\lambda$, denoting the intensity of radiation in the wavelength band between λ and $\lambda + d\lambda$, is given by Plank's famous formula derived under his Quantum Hypothesis and describing the spectral distribution of the radiation from a black body, namely

$$E = \frac{a}{\lambda^5 (e^{b/\lambda} - 1)} , \quad a = bc^3, \quad b = \left(\frac{c}{T}\right) \left(\frac{h}{k}\right) ;$$

here c denotes the velocity of light, T denotes the absolute temperature of the body, k denotes Boltzmann's constant, and h denotes Plank's constant.

Suppose one seeks the distribution law of electromagnetic radiation in the infrared spectrum emitted from a high velocity, high altitude extra terrestrial body following an impact trajectory. As compared with the case of a motionless black body, one observes that such an extra terrestrial body is a grey body (i.e., not a black body or complete radiator) whose shape may be changing quite significantly as it traverses its trajectory to its impact point. Thus at each point of the falling body's trajectory, electromagnetic radiation is being emitted or reflected from a source which is not a black body and whose geometrical configuration can be changing in important detail from point to

point along the body's trajectory.

Let us assume that the trajectory Γ of the body is parameterized as a function $\Gamma = \Gamma(t)$ of time t and denote by $J(t)$ the electromagnetic energy emitted by the body per unit time per unit solid angle about a given direction at time t in the infrared spectrum. As a function of time, we call the radiant intensity $J(t)$ the body's "temporal" distribution of infrared radiation.

There is evidence to suggest that $J(t)$ can be expressed as a function of the body's velocity $v(t)$ and the atmospheric density $\rho(t)$ in the form

$$J(t) = a\rho^b(t)v^c(t)$$

where a , b and c are constants associated with a given class of bodies of similar geometrical shape. Now, provided the temporal distribution law of emitted thermal energy, especially infrared, of a high altitude, high velocity following gray body is of the indicated form, the body's weight w can be expressed as a simple function of the three parameters a , b and c which determine the body's temporal distribution law of electromagnetic radiation in the infrared spectrum. Consequently, if given a time series of observations of the radiant intensity of infrared radiation from a falling gray body, the parameters a , b and c can be accurately estimated, an effective basis for discriminating between several bodies of the same shape falling together according to their size could be achieved.

When the question of determining the temporal distribution law of infrared radiation of a falling gray body from a series of observations of radiant intensity is raised (i.e., estimating the parameters a , b and c in $J(t) = a\rho^b(t)v^c(t)$), another physical model for $J(t)$ may be more appropriate in many situations. This stems from the fact that radio-metric measurements are frequently made under circumstances in which it is impossible to filter the radiation emitted from the body from the effects of the medium through which the radiation has passed or from the radiation reflected from the body from another radiation source. Therefore, under these conditions, one is really dealing not with a model of radiant intensity but with a model of "apparent" radiant intensity, namely radiant intensity having spurious sources of radiation or other spurious variations such as those produced by atmospheric transmission effects.

One such model of "apparent" radiant intensity which combines the spurious sources of radiation or variations, while separating them from the body's own radiant intensity, is given by

$$J^*(t) = a\rho^b(t)v^c(t)\eta(t) = J(t)\eta(t)$$

where $\eta(t)$ is a positive random variable whose probability distribution function is given by a log-normal distribution

$$F(\eta) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_0^\eta \eta^{-1} e^{-\frac{1}{2}\left(\frac{\ln\eta - \mu}{\sigma}\right)^2} d\eta = \Pr(\eta(t) < \eta) .$$

We now combine either of these two physical models of radiant intensity or apparent radiant intensity with a statistical model describing the stochastic behavior of measurement errors which may be present in the observation of any of the four quantities $J(t)$, $J^*(t)$, $\rho(t)$ and $v(t)$. Assuming multiplicative observation errors $\xi(t)$, $\delta(t)$ and $\gamma(t)$ so that

$$J_O(t) = J(t)\xi(t) \quad \text{or} \quad J_O^*(t) = J^*(t)\xi(t)$$

and

$$\rho_O(t) = \rho(t)\delta(t) \quad \text{and} \quad v_O(t) = v(t)\gamma(t) ,$$

we obtain the following stochastic structure representing the physical model and experimental complexes which generate the observation $J_O(t)$ or $J_O^*(t)$ and $\rho_O(t)$ and $v_O(t)$, namely:

$$\ln J(t) = \ln a + b \ln \rho(t) + c \ln v(t)$$

or

$$\ln a + b \ln \rho(t) + c \ln v(t) - \ln J(t) = \ln \eta(t)$$

and

$$\ln J_O(t) = \ln J(t) + \ln \xi(t)$$

$$\ln J_O^*(t) = \ln J^*(t) + \ln \xi(t)$$

$$\ln \rho_O(t) = \ln \rho(t) + \ln \delta(t)$$

$$\ln v_O(t) = \ln v(t) + \ln \gamma(t)$$

where $\eta(t)$, $\xi(t)$, $\delta(t)$ and $\gamma(t)$ may be intercorrelated as well as serially correlated.

In Section 4.2 we see that this stochastic structure is a particular case of a general statistical model which we call the error-in-variable and/or error-in-equation linear structural relationship; and in later sections of this chapter we investigate estimation procedures for determining estimators $\hat{a}$, $\hat{b}$ and $\hat{c}$ of parameters like a , b and c which appear in the above temporal distribution law of electromagnetic energy.

4.2. STOCHASTIC STRUCTURES AND EV OR EVE STATISTICAL MODELS

For our purposes we consider a "stochastic structure" to be a collection whose elements are: (1) a non-empty family of real random variables and a family (possibly null) of real scalar variables; (2) the joint probability distribution or the joint distribution function of these random variables; and (3) a non-empty family of real Baire functions of one or several real variables having the property that compound functions can be formed with each of these Baire functions and the elements of a non-empty subfamily of random or scalar variables of the given families of variables. The scalar variables of a stochastic structure may be scalar quantities, such as means or medians, which are completely determined by the probability distributions of the random variables belonging to the stochastic structure.

Whenever the domain of a Baire function of a stochastic structure contains the cartesian product of the ranges of each member of a non-empty subfamily of random variables, another family of random variables

can be derived by forming the composition of such Baire functions with these subfamilies of random variables. Moreover, the composition of a Baire function jointly with each of the members of a non-empty subfamily of random variables and each of the members of a non-empty subfamily of scalar variables of a stochastic structure can be interpreted as a family of random variables indexed either by those points which belong to the cartesian product of the ranges of the scalar variables or by those points which belong to the cartesian product of the domains of the scalar variables. In either case we call this family of compound random variables, together with their joint probability distribution or joint distribution function, a "derived structure" of the underlying stochastic structure.

Next, let us consider a Baire function whose domain contains the cartesian product of the ranges of each member of a non-null subfamily of random or scalar variables belonging to some stochastic structure. Forming the families of compound functions obtained from this Baire function and this subfamily of random or scalar variables, provided each of its elements is contained in the subfamily of random or scalar variables of the stochastic structure consisting of the complement (in the set-theoretic sense) of the given subfamily of random or scalar variables, we call such a Baire function a "structural relationship" associated with the stochastic structure (through these subfamilies of variables). We note, parenthetically, that any implicit Baire function whose domain contains the cartesian product of the ranges of each member

of a subfamily of random or scalar variables of a stochastic structure yields a structural relation when any one of the variables of the subfamily is expressed as a function of the remaining variables as determined by the implicit Baire function. Finally, the "parameters" of a stochastic structure are the "parameters" of a structural relationship associated with the stochastic structure; and we understand the parameters of a structural relationship to be any unknown (constant or variable) scalars whose specification determine the Baire function of the structural relation.

We now restate this verbal presentation of the notion of a stochastic structure, its derived structure, and a structural relationship more conveniently and more succinctly in terms of mathematical notation and description. Our basic mathematical structure consists of

$\{(\Omega_1, A_1), (\Omega_2, A_2), \dots, (\Omega_N, A_N)\}$, a finite family of measurable spaces;

$\{Y_1, Y_2, \dots, Y_N\}$, a finite family of random variables; and

$\{P_1, P_2, \dots, P_N\}$, a finite family of probability measures,

each of whose members has as its domain the similarly indexed σ -field of measurable sets. Each of the members of the given family of random variables, with values in the real Borel measurable space (R, B) , is defined on the correspondingly indexed measurable space and its probability distribution is the identically indexed probability measure. In addition we have

$\{y_1, y_2, \dots, y_M\}$, a finite family of scalar variables,

each of whose members is a function with domain and range space R (the real line).

As previously indicated, we are interested in the joint probability distribution P of the family of random variables rather than the family of individual probability distributions of these random variables. The probability measure P has as its domain the σ -field product A of the σ -field $A_1, A_2, \dots$, and A_N ; and it may be interpreted as the probability distribution of the random vector $V_N = (Y_1, Y_2, \dots, Y_N)$ which takes its values in the N -dimensional real Borel measurable space (R^N, B^N) and is defined on the measurable space (Ω, A) where Ω is the cartesian product of the N spaces $\Omega_1, \Omega_2, \dots$, and Ω_N .

Corresponding to the family of individual probability distributions $\{P_k\}_{k=1}^N$ is the family of individual distribution function $\{F_k(v_k)\}_{k=1}^N$ where

$$F_k(v_k) = P_k(Y_k^{-1}(-\infty, v_k)) ;$$

or, corresponding to the joint probability distribution P of the family $\{Y_k\}_{k=1}^N$ of random variables is their joint distribution function

$$F(v_1, v_2, \dots, v_N) = P(V_N^{-1} \bigcap_{k=1}^N (-\infty, v_k)) .$$

The family of random variables $\{Y_k\}_{k=1}^N$ and the family of scalar variables $\{y_k\}_{k=1}^M$, together with the joint probability distribution P or the joint distribution function F , comprise basic elements of what we have called a stochastic structure. We need only specify in addition,

$$\{\phi_1, \phi_2, \dots, \phi_H\}, \text{ a family of Baire functions,}$$

for which each member ϕ_k of this family has as its domain the H_k -dimen-

sional Borel measurable space (R^{H_k}, B^{H_k}) and as its range the one-dimensional Borel measurable space (R, B) .

Now the set of scalar variables $\{y_k\}_{k=1}^M$, the set of random variables $\{Y_k\}_{k=1}^N$, the set of Baire functions $\{\phi_k\}_{k=1}^H$, and the joint probability distribution P or the joint distribution function F when considered jointly as elements of an abstract collection from a stochastic structure which we denote by,

$$S = \left[\{y_k\}_{k=1}^M, \{Y_k\}_{k=1}^N, \{\phi_k\}_{k=1}^H, P \right],$$

where P may be replaced by F . We understand, of course, that through composition of the Baire function

$$x_k = \phi_k(z_1, z_2, \dots, z_{p_k}, z_{p_k+1}, \dots, z_{p_k+q_k})$$

from S with $z_i = Y_{m_i}$ and $z_{p_k+j} = y_{n_j}$ for $1 \leq i \leq p_k$, $1 \leq j \leq q_k$, $p_k+q_k = H_k$, we obtain the random variable

$$X_k^{(y)} = \phi_k(Y_{m_1}, Y_{m_2}, \dots, Y_{m_p}, y_{n_1}, y_{n_2}, \dots, y_{n_q})$$

or since $y_k = y_k(t_k)$,

$$X_k^{(t)} = \phi_k(Y_{m_1}, Y_{m_2}, \dots, Y_{m_p}, y_{n_1}(t_1), y_{n_2}(t_2), \dots, y_{n_q}(t_q))$$

for each fixed but arbitrary index

$$y = (y_{n_1}, y_{n_2}, \dots, y_{n_q}) \quad \text{or} \quad t = (t_1, t_2, \dots, t_q).$$

Here $\{Y_{m_i}\}_{i=1}^{P_k}$ and $\{y_{n_j}\}_{j=1}^{q_k}$ are specified subfamilies of random or scalar variables, respectively, from S .

The random variable $X_k^{(y)}$ or $X_k^{(t)}$ is defined on the measurable space $(\Omega^{(N_k)}, A^{(N_k)})$ with values in (R, B) , where $N_k = \{m_1, m_2, \dots, m_{p_k}\}$ and $\Omega^{(N_k)}$ is the cartesian product of the spaces $\Omega_{m_1}, \Omega_{m_2}, \dots$, and $\Omega_{m_{p_k}}$ and $A^{(N_k)}$ is the σ -field product of the σ -fields $A_{m_1}, A_{m_2}, \dots$, and $A_{m_{p_k}}$. Each random variable $X_k^{(y)}$ is distributed by a probability measure $Q_k^{(y)}$ related to the joint probability distributions $P_{(N_k)}$ of the family of random variables $\{Y_{m_i}\}_{i=1}^{P_k}$. In turn, the random vector

$$U_H^{(y)} = (X_1^{(y)}, X_2^{(y)}, \dots, X_H^{(y)})$$

is defined on (Ω^*, A^*) with values in (R^H, B^H) where Ω^* is the cartesian product of the H spaces $\Omega^{(N_k)}$ and A^* is the σ -field product of the H σ -fields $A^{(N_k)}$. Moreover, $U_H^{(y)}$ is distributed by a probability measure $Q^{(y)}$ which is the joint probability distribution of the family of random variables $\{X_k^{(y)}\}_{k=1}^H$. Thus $Q^{(y)}$ is related to the probability distributions of the family $\{Q_k^{(y)}\}_{k=1}^H$ and consequently to the family of joint probability distributions $\{P_{(N_k)}\}_{k=1}^H$ of the H subfamilies of random variables $\{Y_{m_i}\}_{i=1}^{P_k}$.

Finally, besides the probability measures $\{Q_k^{(y)}\}_{k=1}^H$ or $Q^{(y)}$, we also consider the family of distribution functions $\{G_k^{(y)}(u_k)\}_{k=1}^H$ of the family of random variables $\{X_k^{(y)}\}_{k=1}^H$ where,

$$G_k^{(y)}(u_k) = Q_k^{(y)}(X_k^{-1}(-\infty, u_k)) ;$$

or, the joint distribution function,

$$G^{(y)}(u_1, u_2, \dots, u_H) = Q^{(y)}((U_H^{(y)})^{-1} \bigtimes_{k=1}^H (-\infty, u_k)) ,$$

of the random vector $U_H^{(y)} = (X_1^{(y)}, X_2^{(y)}, \dots, X_H^{(y)})$.

Keeping these constructions in mind, we see that the derived structure of the stochastic structure S is the collection whose elements are the family of random variables $\{X_k^{(y)}\}_{k=1}^H$, the family of scalar variables $\{y_{(m_k)}\}_{k=1}^H$ where $y_{(m_k)} = \{y_{n_j}\}_{j=1}^{q_k}$, and the joint probability distribution $Q^{(y)}$ or the joint distribution function $G^{(y)}(u_1, u_2, \dots, u_H)$. We denote the derived structure of S by,

$$D[S] = \left[\{ \{y_{u_j}\}_{j=1}^{q_k} \}_{k=1}^H, \{X_k^{(y)}\}_{k=1}^H, Q^{(y)} \right] ,$$

where $Q^{(y)}$ may be replaced by $G^{(y)}$.

Next, let us consider another Baire function on (R^r, B^r) to (R, B) which depends on a parameter $\theta = (\theta_1, \theta_2, \dots, \theta_s)$ in E^s , say

$$z = f(z_1, z_2, \dots, z_r; \theta_1, \theta_2, \dots, \theta_s) ;$$

or, briefly, $z = f(z_1, z_2, \dots, z_r; \theta)$. If the family of scalar variables of the stochastic structure contains a subfamily $y_{j_1}, y_{j_2}, \dots, y_{j_r}$ of scalar variables so that

$$y = f(y_{j_1}, y_{j_2}, \dots, y_{j_r}; \theta)$$

belongs to the subfamily $\{y_j\}_{j=1}^M - \{y_{j_q}\}_{q=1}^r$ of S , we say that f is a structural relation associated with the stochastic structure S through

this subfamily of scalar variables. We also call f a stochastic structure associated with S whenever subfamilies $\{Y_{i_1}, Y_{i_2}, \dots, Y_{i_u}\}$ of random variables or $\{y_{j_1}, y_{j_2}, \dots, y_{j_v}\}$ of scalar variables of S exist, where $r = u+v$, for which the random variables,

$$Y(y) = f(Y_{i_1}, Y_{i_2}, \dots, Y_{i_u}, y_{j_1}, y_{j_2}, \dots, y_{j_v}; \theta)$$

belong to the subfamily $\{Y_i\}_{i=1}^N - \{Y_{i_p}\}_{p=1}^u$ of S . In either case θ is called a parameter of this structural relationship or a parameter of the stochastic structure.

We use the notation (f, S_f) to denote that f is a structural relation associated with the stochastic structure S . Frequently, we replace S_f by $D[S_f]$ in this notation in order to indicate that although f is a structural relation associated with S , in the particular context in which the notation $(f, D[S_f])$ appears, it is f and the derived structure of S that are of primary concern. In many instances, for example, only the random variables of a derived structure of an underlying stochastic structure are observable. Consequently, in these circumstances, estimates of the parameters of a structural relationship associated with a stochastic structure can be obtained only through functions of observations of the random variables of the derived structure. We see that this is the case in the EV and EVE statistical models which we introduce now as particular stochastic structures and specific structural relationships associated with them.

Consider the stochastic structure,

$$S = \left[\{r_i(t)\}_{i=1}^p, \{\delta_i(t)\}_{i=1}^p, \phi(z_1, z_2) = z_1 + z_2 \right],$$

where the $r_i(t)$ are scalar functions and the $\delta_i(t)$ are random variables having $E\delta_i(t) = 0$ and whose distribution function

$$F_t(\delta_1, \delta_2, \dots, \delta_p) = \Pr(\delta_1(t) < \delta_1, \delta_2(t) < \delta_2, \dots, \delta_p(t) < \delta_p)$$

remains unspecified except that the $\{\delta_i(t)\}_{i=1}^p$ are assumed to have the known covariance matrix

$$\Sigma_\delta(t, t) = E(\delta^T(t)\delta(t))$$

where $\delta(t) = (\delta_1(t), \delta_2(t), \dots, \delta_p(t))$. From S we obtain the derived structure

$$D[S] = \left[\{r_i(t)\}_{i=1}^p, \{X_i(t)\}_{i=1}^p \right]$$

where, for $1 \leq i \leq p$,

$$X_i(t) = \phi(r_i(t), \delta_i(t)) = r_i(t) + \delta_i(t).$$

Given S , the function

$$f(z_1, z_2, \dots, z_p; \alpha_1, \alpha_2, \dots, \alpha_p) = \sum_{i=1}^p \alpha_i z_i = 0$$

when compounded with $z_1 = r_1(t), z_2 = r_2(t), \dots$, and $z_p = r_p(t)$ leads to the structural relationship

$$\alpha_1 r_1(t) + \alpha_2 r_2(t) + \dots + \alpha_p r_p(t) = 0$$

associated with the stochastic structure S or its derived structure $D[S]$ through the family $\{r_i(t)\}_{i=1}^p$ of scalar variables. We call (f, S_f) or $(f, D[S_f])$ for this particular stochastic structure together with this linear structural relationship the "error-in-variable" statistical model or merely the EV statistical model.

Assuming that the variables $r_i(t)$ of (f, S_f) are not directly observable due to the "errors-in-variables" $\delta_i(t)$, but that only the $X_i(t)$ of the derived structure $D[S_f]$ are observable, we wish to estimate the unknown parameters of f from a sample of n -observations of $\{X_i(t)\}_{i=1}^p$ at, say, $t = t_1, t = t_2, \dots$, and $t = t_n$. In other words, given the known covariance matrices

$$\Sigma_X(t_i, t_j) = E(X^T(t_i)X(t_j)) = E(\delta^T(t_i)\delta(t_j)) = \Sigma_\delta(t_i, t_j)$$

among the errors we seek estimators

$$\hat{\alpha}_k = \hat{\alpha}_k(X_1(t_1), X_2(t_1), \dots, X_p(t_1), \dots, X_1(t_n), X_2(t_n), \dots, X_p(t_n))$$

of the unknown parameters α_k of the linear structural relationship $\sum_{i=1}^p \alpha_i r_i(t) = 0$ associated with the stochastic structure S or the EV statistical model.

Another important stochastic structure is given by $S_* =$

$$\left[\{r_i(t)\}_{i=1}^p, \{S_j(t)\}_{j=1}^q, \xi(t), \{\delta_i(t)\}_{i=1}^p, \{\eta_j(t)\}_{j=1}^q, \phi(z_1, z_2) = z_1 + z_2 \right]$$

where the $r_i(t)$'s are scalar variables and the $\xi(t)$'s, $\delta_i(t)$'s, $\eta_j(t)$'s and $S_j(t)$'s are random variables for which $E(\xi(t)) = 0$, $E(\delta_i(t)) = 0$ and $E(\eta_j(t)) = 0$; and $z = \phi(z_1, z_2) = z_1 + z_2$ is understood to be composed with $z_1 = r_i(t)$ and $z_2 = \delta_i(t)$ for $1 \leq i \leq p$ and with $z_1 = S_j(t)$ and $z_2 = \eta_j(t)$ for $1 \leq j \leq q$.

Associated with S_* is the derived structure

$$D[S_*] = \left[\{r_i(t)\}_{i=1}^p, \{X_i(t)\}_{i=1}^p, \{Y_j(t)\}_{j=1}^q \right]$$

where, for $1 \leq i \leq p$,

$$X_i(t) = \phi(r_i(t), \delta_i(t)) = r_i(t) + \delta_i(t) ,$$

and, for $1 \leq j \leq q$,

$$Y_j(t) = \phi(S_j(t), \eta_j(t)) = S_j(t) + \eta_j(t) .$$

Next, given S and the function

$$z_0 = \sum_{i=1}^p \alpha_i z_i + \sum_{j=1}^q \beta_j z_{j+p} = f_*(z_1, z_2, \dots, z_{p+q}; \alpha, \beta) ,$$

taking

$$z_0 = \xi(t), z_1 = r_1(t), \dots, z_p = r_p(t), z_{p+1} = S_1(t), \dots, \text{ and } z_{p+q} = S_q(t)$$

and forming through f_* the compound function

$$\sum_{i=1}^p \alpha_i r_i(t) + \sum_{j=1}^q \beta_j S_j(t) = \xi(t)$$

we obtain a linear structural relation associated with S_* or $D[S_*]$ through the indicated families of scalar and random variables.

Finally, putting

$$\gamma(t) = (\xi(t), \delta(t), \eta(t))$$

where

$$\delta(t) = (\delta_1(t), \delta_2(t), \dots, \delta_p(t))$$

and

$$\eta(t) = (\eta_1(t), \eta_2(t), \dots, \eta_q(t)) ,$$

we assume that the random vector $\gamma(t)$ has a known covariance matrix

$$\Sigma_\gamma(t, t) = \begin{bmatrix} \sigma_\xi(t, t) & \Sigma_{(\xi, \delta)}(t, t) & \Sigma_{(\xi, \eta)}(t, t) \\ \Sigma_{(\delta, \xi)}(t, t) & \Sigma_\delta(t, t) & \Sigma_{(\delta, \eta)}(t, t) \\ \Sigma_{(\eta, \xi)}(t, t) & \Sigma_{(\eta, \delta)}(t, t) & \Sigma_\eta(t, t) \end{bmatrix} = E(\gamma^T(t) \gamma(t))$$

$$\begin{aligned} \text{where } \sigma_\xi(t, t) &= E\xi^2(t), \Sigma_{(\xi, \delta)}(t, t) = E(\xi(t)\delta(t)), \Sigma_{(\xi, \eta)}(t, t) \\ &= E(\xi(t)\eta(t)), \Sigma_{(\delta, \xi)}(t, t) = E(\delta^T(t)\xi(t)), \Sigma_\delta(t, t) = E(\delta^T(t)\delta(t)), \text{ etc.} \end{aligned}$$

We call (f_*, S_*) or $(f_*, D[S_*])$ for the particular stochastic structure S_* or the derived structure $D[S_*]$ together with the linear structural relationship f_* the "error-in-variable and/or error-in-equation" statistical model or simply the EVE statistical model. When $r_i(t)$ and $S_j(t)$ are not directly observable, having errors $\delta_i(t)$ and $\eta_j(t)$, and

only the random variables $\{X_i(t)\}_{i=1}^p$ and $\{Y_j(t)\}_{j=1}^q$ of the derived structure of the EVE statistical model are observable, we seek estimators

$$\hat{\alpha}_i = \hat{\alpha}_i(X_1(t_1), X_2(t_1), \dots, X_p(t_1), \dots, Y_1(t_n), Y_2(t_n), \dots, Y_q(t_n))$$

and

$$\hat{\beta}_j = \hat{\beta}_j(X_1(t_1), X_2(t_1), \dots, X_p(t_1), \dots, Y_1(t_n), Y_2(t_n), \dots, Y_q(t_n)) .$$

for α_i and β_j , $1 \leq i \leq p$ and $1 \leq j \leq q$, given a sample of n -observations of $\{X_i(t)\}_{i=1}^p$ and $\{Y_j(t)\}_{j=1}^q$ at $t = t_1, t = t_2, \dots$, and $t = t_n$ when the errors $\gamma(t) = (\xi(t), \delta(t), \eta(t))$ can be intercorrelated and serially correlated with known covariance matrices given by

$$\Sigma_\gamma(t_i, t_j) = E(\gamma^T(t_i) \gamma(t_j)) .$$

In the following sections of this chapter we discuss the method of generalized least squares estimation for estimating the unknown coefficient parameters of the EV and EVE statistical models when the variable and equation errors are both intercorrelated and serially correlated. We derive the non-linear system of normal equations which implicitly define these generalized least squares estimators by using the Dwyer-Macphail matrix derivative methods developed in the previous two chapters. Besides illustrating the theoretical aspects of the Dwyer-Macphail matrix derivative methods in terms of a concrete example, this application also points out the simplicity and ease of application of these matrix differentiation techniques once the fundamental rules of this matrix derivative calculus are learned.

4.3. GENERALIZED LEAST SQUARES ESTIMATORS FOR EV MODEL

We begin our discussion of the estimation problem in the EV model by first defining what we mean when we speak of the "generalized least squares" estimation procedure. For this purpose, let U_{θ} be a random vector in Φ^n which is indexed by an (unknown) vector parameter θ in Φ^s , and suppose that $EU_{\theta} = 0$ while the covariance matrix $\Sigma_U = EU_{\theta}^T U_{\theta}$ of U_{θ} is a positive definite matrix for any θ in a specified domain Θ in Φ^s . Under these circumstances, we call θ_0 a "generalized least squares" estimator of the parameter θ whenever θ_0 is an extreme point in Θ of the metric $\|U_{\theta}\|_W = U_{\theta}^T W U_{\theta}$ (considered as a function of θ) determined by the matrix weight $W = \Sigma_U^{-1}$ for which

$$\|U_{\theta_0}\|_W = \min \{ \|U_{\theta}\|_W \mid \theta \in \Theta \} .$$

Having defined what we mean by the "generalized least squares" estimation procedure and "generalized least squares" estimators of an unknown parameter θ , which we refer to merely as the GLS-estimation procedure or GLS-estimators, we go on to derive GLS-estimators of the unknown parameters of the EV statistical model. For this purpose, let us write the EV model in vector form and adapt some other convenient notation. Thus let

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p) ,$$

$$r(t) = (r_1(t), r_2(t), \dots, r_p(t)) ,$$

and

$$X(t) = (X_1(t), X_2(t), \dots, X_p(t)) = r(t) + \delta(t) ;$$

then in this vector notation we may write the EV statistical model simply as

$$\alpha \cdot r^T(t) = 0 \quad \text{where} \quad X(t) = r(t) + \delta(t)$$

for $t = t_1, t = t_2, \dots$, and $t = t_n$. Finally, rather than writing, for example, $X(t_1), X(t_2)$, etc., we merely write $X(1), X(2)$, etc.; and we denote $\sum \delta(t_i, t_j)$ simply as $\sum_{ij}$.

Now then, how do we obtain the GLS-estimators of the **parameters** for the EV statistical model when the "errors" $\delta(t)$ are known to be inter-correlated and serially correlated. In this connection, let us consider the "deviation,"

$$d_i(\alpha) = \alpha \cdot X^T(i) ,$$

corresponding to the i -th observation $X(i)$ for $1 \leq i \leq n$; or, alternately, the vector of deviations ,

$$d(\alpha) = (d_1(\alpha), d_2(\alpha), \dots, d_n(\alpha)) ,$$

corresponding to the n observations $\{X(i)\}_{i=1}^n$ jointly. We easily see that

$$E(d(\alpha)) = 0$$

and that

$$\Sigma_d(\alpha) = [\sigma_{ij}(\alpha)] = E(d^T(\alpha)d(\alpha))$$

where

$$\sigma_{ij}(\alpha) = \alpha \sum_{i,j} \alpha^T \quad \text{for } 1 \leq i, j \leq n.$$

Writing $W(\alpha) = \Sigma_d^{-1}(\alpha)$ and assuming that $W(\alpha)$ exists whenever α belongs to an appropriate domain in Φ^P , we are led to the generalized least squares estimation criterion

$$Q(\alpha) = d(\alpha)W(\alpha)d^T(\alpha)$$

for the unknown coefficient parameters of the EV statistical model.

Thus the application of the GLS-estimation procedure to the EV model leads us to the problem of finding critical points of a scalar function of a vector variable which is the (matrix) product of vector and matrix functions of a vector variable. Therefore, adopting the generalized least squares estimation procedure, the estimation problem in the EV statistical model may be simply and succinctly stated as follows:

Using the Dwyer-Macphail matrix derivative methods, obtain necessary conditions for a critical point of the functional of a vector variable given by

$$Q(\alpha) = \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) w_{ij}(\alpha) X(j) \alpha^T = d(\alpha)W(\alpha)d^T(\alpha)$$

where

$$W(\alpha) = [w_{ij}(\alpha)] = [\alpha \sum_{i,j} \alpha^T]^{-1} = \Sigma_d^{-1}(\alpha).$$

Then a GLS-estimator $\hat{\alpha}$ of α is a critical point of $Q(\alpha)$ which is a local minimum.

In order to derive the non-linear normal equations which define this GLS-estimator $\hat{\alpha}$ of the unknown parameter α of the EV statistical model, we rewrite $Q(\alpha)$ as a sum of products of scalar functions of a vector variable, namely

$$Q(\alpha) = \sum_{i=1}^n \sum_{j=1}^n f_{ij}(\alpha) h_{ij}(\alpha)$$

where $f_{ij}(\alpha) = w_{ij}(\alpha)$ and $h_{ij}(\alpha) = \alpha X^T(i) X(j) \alpha^T$.

Now we find that

$$\frac{\partial Q(\alpha)}{\partial \langle \alpha \rangle_k} = \sum_{i=1}^n \sum_{j=1}^n \left(\frac{\partial f_{ij}(\alpha)}{\partial \langle \alpha \rangle_k} h_{ij}(\alpha) + f_{ij}(\alpha) \frac{\partial h_{ij}(\alpha)}{\partial \langle \alpha \rangle_k} \right) ;$$

and, according to Theorem 2.4.2,

$$\frac{\partial Q(\alpha)}{\partial \alpha} = \left[\text{Tr} \left(\frac{\partial Q(\alpha)}{\partial \langle \alpha \rangle_k} K^T \right) \right]_k .$$

But since $Q(\alpha)$ is a scalar function, the matrix K appearing in this expression is merely the scalar constant $K = 1$ and, therefore,

$$\text{Tr} \left(\frac{\partial Q(\alpha)}{\partial \langle \alpha \rangle_k} K^T \right) = \frac{\partial Q(\alpha)}{\partial \langle \alpha \rangle_k} .$$

Consequently,

$$\frac{\partial Q(\alpha)}{\partial \alpha} = \sum_{i=1}^n \sum_{j=1}^n h_{ij}(\alpha) \left[\frac{\partial f_{ij}(\alpha)}{\partial \langle \alpha \rangle_k} \right]_k + f_{ij}(\alpha) \left[\frac{\partial h_{ij}(\alpha)}{\partial \langle \alpha \rangle_k} \right]_k ;$$

or, equivalently, since

$$\frac{\partial f_{ij}(\alpha)}{\partial \alpha} = \left[\frac{\partial f_{ij}(\alpha)}{\partial \langle \alpha \rangle_k} \right]_k \quad \text{and} \quad \frac{\partial h_{ij}(\alpha)}{\partial \alpha} = \left[\frac{\partial h_{ij}(\alpha)}{\partial \langle \alpha \rangle_k} \right]_k,$$

the Dwyer-Macphail matrix derivative of $Q(\alpha)$ is given by,

$$\frac{\partial Q(\alpha)}{\partial \alpha} = \sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{\partial f_{ij}}{\partial \alpha} + f_{ij} \frac{\partial h_{ij}}{\partial \alpha},$$

where $\partial f_{ij}/\partial \alpha$ is the Dwyer-Macphail matrix derivative of $f_{ij} = w_{ij}(\alpha)$

and $\partial h_{ij}/\partial \alpha$ is the Dwyer-Macphail matrix derivative of

$$h_{ij} = \alpha X^T(i)X(j)\alpha^T.$$

Thus to determine $\partial Q(\alpha)/\partial \alpha$ we need only find $\partial h_{ij}/\partial \alpha$ and $\partial f_{ij}/\partial \alpha$.

We see immediately that

$$\begin{aligned} \frac{\partial h_{ij}}{\partial \langle \alpha \rangle_k} &= \frac{\partial}{\partial \langle \alpha \rangle_k} (\alpha X^T(i)X(j)\alpha^T) \\ &= J_k X^T(i)X(j)\alpha^T + \alpha X^T(i)X(j) J_K^T; \end{aligned}$$

and, therefore, according to Theorem 2.4.3, noting that $K = 1 = K^T$ since $h_{ij}(\alpha)$ is a scalar function, we obtain the Dwyer-Macphail matrix derivative of $h_{ij}(\alpha)$ as

$$\begin{aligned} \frac{\partial h_{ij}}{\partial \alpha} &= 1 \cdot \alpha X^T(j)X(i) + 1 \cdot \alpha X^T(i)X(j) \\ &= \alpha (X^T(i)X(j) + X^T(j)X(i)). \end{aligned}$$

Next, we turn our attention to deriving the Dwyer-Macphail matrix derivative of $f_{ij} = w_{ij}(\alpha)$. Writing,

$$U = U(Y) = Y^{-1} \quad \text{and} \quad Y = V(\alpha) = \Sigma_d(\alpha) ,$$

we see that

$$W = U(V(\alpha)) = \Sigma_d^{-1}(\alpha) = W(\alpha) .$$

Consequently, through an application of Theorem 2.4.4, we can write the Dwyer-Macphail matrix derivative of $w_{ij}(\alpha) = \langle W(\alpha) \rangle_{ij}$ in terms of the Dwyer-Macphail matrix derivatives of $U = Y^{-1}$ and $Y = \Sigma_d(\alpha)$ as follows:

$$\frac{\partial w_{ij}(\alpha)}{\partial \alpha} = \sum_{r=1}^n \sum_{s=1}^n \left\langle \frac{\partial \langle U \rangle_{ij}}{\partial Y} \right\rangle_{rs} \frac{\partial \langle Y \rangle_{rs}}{\partial \alpha} .$$

Since $\langle Y \rangle_{rs} = \langle \Sigma_d(\alpha) \rangle_{rs} = \alpha \Sigma_{rs} \alpha^T$, we find that

$$\begin{aligned} \frac{\partial \langle Y \rangle_{rs}}{\partial \langle \alpha \rangle_k} &= \frac{\partial}{\partial \langle \alpha \rangle_k} (\alpha \Sigma_{rs} \alpha^T) \\ &= J_k \Sigma_{rs} \alpha^T + \alpha \Sigma_{rs} J_k^T ; \end{aligned}$$

and thus, since $\langle Y \rangle_{rs} = \alpha \Sigma_{rs} \alpha^T$ is a scalar and therefore $K = 1 = K^T$,

$$\begin{aligned} \frac{\partial \langle Y \rangle_{rs}}{\partial \alpha} &= 1 \cdot \alpha \Sigma_{rs}^T + 1 \cdot \alpha \Sigma_{rs} \\ &= \alpha (\Sigma_{rs} + \Sigma_{sr}) \\ &= 2\alpha \Sigma_{rs} \end{aligned}$$

where $\Sigma_{rs}^T = (E \delta^T(r) \delta(s))^T = E(\delta^T(r) \delta(s))^T = \Sigma_{sr}$ and $\Sigma_{rs} = \Sigma_{sr}$ because $\Sigma_d(\alpha) = [\alpha \Sigma_{ij} \alpha^T]$ is a covariance matrix and must be symmetric.

Without repeating other well known derivations, but referring the interest reader to [32] for example, we write down the Dwyer-Macphail

matrix derivative of $\langle U \rangle_{ij} = \langle Y^{-1} \rangle_{ij}$ as

$$\frac{\partial \langle U \rangle_{ij}}{\partial Y} = -U^T K_{ij} U^T .$$

Recalling for a matrix A that $(A)_k$ denotes its k-th row, we easily find that

$$\left\langle \frac{\partial \langle U \rangle_{ij}}{\partial Y} \right\rangle_{rs} = - \langle (U)_i^T (U^T)_j \rangle_{rs} .$$

Hence, combining these results into a single expression for the Dwyer-Macphail matrix derivative of $f_{ij} = w_{ij}(\alpha)$ according to the "chain" rule for matrix differentiation of a compound function given by Theorem 2.4.4, we see that

$$\frac{\partial w_{ij}(\alpha)}{\partial \alpha} = -2\alpha \left(\sum_{r=1}^n \sum_{s=1}^n \langle (W(\alpha))_i^T (W(\alpha))_j \rangle_{rs} \Sigma_{rs} \right) .$$

Furthermore, having derived the Dwyer-Macphail matrix derivatives of $f_{ij} = w_{ij}(\alpha)$ and $h_{ij} = \alpha X^T(i) X(j) \alpha^T$, we can now give an explicit expression for the Dwyer-Macphail matrix derivative of $Q(\alpha)$, namely:

$$\begin{aligned} \frac{\partial Q(\alpha)}{\partial \alpha} = & 2\alpha \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij}(\alpha) \left(\frac{X^T(i) X(j) + X^T(j) X(i)}{2} \right) \right) \\ & - 2\alpha \left(\sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) X(j) \alpha^T \left(\sum_{r=1}^n \sum_{s=1}^n \langle (W(\alpha))_i^T (W(\alpha))_j \rangle_{rs} \Sigma_{rs} \right) \right) . \end{aligned}$$

This expression for $\partial Q(\alpha)/\partial \alpha$ can be further refined by noting that for any square matrix W in $\Phi_{(n,n)}$,

$$(W^T)_r K_{ij} (W^T)_s = \langle (W)_i^T (W^T)_j \rangle_{rs} ,$$

where $\{W^T\}_s$ denotes the s -th column of W^T . Using this identity we see that

$$\begin{aligned}
& \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) X(j) \alpha^T \left(\sum_{r=1}^n \sum_{s=1}^n \langle (W(\alpha))_i^T (W(\alpha))_j \rangle_{rs} \right) \Sigma_{rs} \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) X(j) \alpha^T \left(\sum_{r=1}^n \sum_{s=1}^n (W(\alpha))_r K_{ij} (W(\alpha))_s \right) \Sigma_{rs} \\
&= \sum_{r=1}^n \sum_{s=1}^n \left(\sum_{i=1}^n \sum_{j=1}^n (W(\alpha))_r (\alpha X^T(i) X(j) \alpha^T K_{ij}) (W(\alpha))_s \right) \Sigma_{rs} \\
&= \sum_{r=1}^n \sum_{s=1}^n (W(\alpha))_r [\alpha X^T(i) X(j) \alpha^T]_{(i,j)} (W(\alpha))_s \Sigma_{rs}
\end{aligned}$$

In turn we see that

$$\begin{aligned}
& (W(\alpha))_r [\alpha X^T(i) X(j) \alpha^T] (W(\alpha))_s \\
&= \text{Tr} \left((W(\alpha))_r [\alpha X^T(i) X(j) \alpha^T] (W(\alpha))_s \right) \\
&= \text{Tr} \left([\alpha X^T(i) X(j) \alpha^T] [(W(\alpha))_r^T (W(\alpha))_s^T] \right) \\
&= \text{Tr} \left([X^T \alpha^T] [\alpha X] [(W(\alpha))_r^T (W(\alpha))_s^T] \right) \\
&= \text{Tr} \left([\alpha X] [(W(\alpha))_r^T (W(\alpha))_s^T] [X^T \alpha^T] \right) \\
&= [\alpha X] [(W(\alpha))_r^T (W(\alpha))_s^T] [X^T \alpha^T] \\
&= (\alpha X (W(\alpha))_r^T) ((W(\alpha))_s^T X^T \alpha^T) \\
&= (\alpha X (W(\alpha))_r) ((W(\alpha))_s^T X^T \alpha^T) \\
&= (\alpha X W(\alpha) e_r^T) (e_s^T W^T(\alpha) X^T \alpha^T) \\
&= \mu_r(\alpha) \cdot \mu_s(\alpha)
\end{aligned}$$

where it is understood that e_k is the row vector of order n having one in its k -th place and zero elsewhere,

$$X = [X^T(1), X^T(2), \dots, X^T(n)] ,$$

and

$$\begin{aligned} \mu_k(\alpha) &= \alpha X W(\alpha) e_k^T \\ &= (\alpha X \{W(\alpha)\}_1, \alpha X \{W(\alpha)\}_2, \dots, \alpha X \{W(\alpha)\}_n) e_k^T \\ &= \alpha X \{W(\alpha)\}_k . \end{aligned}$$

Combining these results we see that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) X(j) \alpha^T \left(\sum_{r=1}^n \sum_{s=1}^n \langle (W(\alpha))_i^T (W(\alpha))_j \rangle_{rs} \Sigma_{rs} \right) \\ = \sum_{r=1}^n \sum_{s=1}^n \mu_r(\alpha) \Sigma_{rs} \mu_s(\alpha) ; \end{aligned}$$

and, therefore, that

$$\frac{\partial Q(\alpha)}{\partial \alpha} = 2\alpha \left(\sum_{r=1}^n \sum_{s=1}^n w_{rs}(\alpha) \left(\frac{X^T(r) X(s) + X^T(s) X(r)}{2} \right) - \mu_r(\alpha) \Sigma_{rs} \mu_s(\alpha) \right) .$$

Keeping this derivation of $\partial Q(\alpha)/\partial \alpha$ in mind, we now state

Theorem 4.3.1. Suppose the stochastic structure (f, S) is the EV statistical model where f is given implicitly by the linear structural relationship $\alpha \cdot r^T(i) = 0$ and where only the random (vector) variables $X(i) = r(i) + \delta(i)$ of the derived structure $D[S]$ of S are observable for

$1 \leq i \leq n$, while $r(i)$ is non-observable having "errors" $\delta(i)$ with $E\delta(i) = 0$. Let $d(\alpha)$ be the vector whose i -th component is $d_i(\alpha) = \alpha \cdot X^T(i)$ and assume that

$$W(\alpha) = [w_{ij}(\alpha)] = [\alpha \sum_{ij} \alpha^T]^{-1} = \Sigma_d^{-1}(\alpha)$$

exists for every non-zero α where $\sum_{ij} = E(\delta^T(i)\delta(j))$ for $1 \leq i, j \leq n$ are known error covariance matrices. Then any non-trivial generalized least squares estimator $\hat{\alpha}$ of α must be a solution of the non-linear equation system

$$A_\alpha \alpha^T = 0$$

where

$$A_\alpha = \sum_{r=1}^n \sum_{s=1}^n \left(w_{rs}(\alpha) \left(\frac{X^T(r)X(s) + X^T(s)X(r)}{2} \right) - \mu_r(\alpha) \sum_{rs} \mu_s(\alpha) \right)$$

Here $\mu_k(\alpha) = \alpha X W(\alpha) e_k^T$ where X is the matrix whose i -th column is the vector $X^T(i)$ for $1 \leq i \leq n$ and e_k is the n -dimensional row vector having one in its k -th place and zero elsewhere.

Proof. According to Theorem 3.5.4, if α_0 is any non-zero critical point of the GLS criterion

$$Q(\alpha) = \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) w_{ij}(\alpha) X(j) \alpha^T = d(\alpha) W(\alpha) d^T(\alpha)$$

then the Dwyer-Macphail matrix derivative $\partial Q / \partial \alpha_0$ must vanish. But we have seen that

$$\frac{\partial Q(\alpha)}{\partial \alpha} = 2\alpha \sum_{r=1}^n \sum_{s=1}^n \left(w_{ij}(\alpha) \left(\frac{X^T(r)X(s) + X^T(s)X(r)}{2} \right) - \mu_r(\alpha) \sum_{rs} \mu_s(\alpha) \right) ;$$

consequently, if α_0 is a non-zero critical point of $Q(\alpha)$, then α_0 must satisfy the non-linear system of equations

$$A_Q \alpha^T = 0 .$$

In particular, since $\hat{\alpha}$ is a non-trivial generalized least squares estimator of α , $\hat{\alpha}$ is a critical point of $Q(\alpha)$ and therefore must satisfy this non-linear system of equations. Thus the proof of Theorem 4.3.1 is completed.

Before taking up our discussion of the estimation problem in the EVE statistical model, we explore one more facet of the estimation problem for the EV model, namely the adaptation of the GLS-estimation procedure to an EV statistical model in which the structural relationship is given explicitly, rather than implicitly, as

$$r(i) = \alpha_1 + \alpha_2 r_2(i) + \dots + \alpha_{p-1} r_{p-1}(i) .$$

We find, of course, that the GLS-estimator $(\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{p-1})$ of $(\alpha_1, \alpha_2, \dots, \alpha_{p-1})$ is defined by a different system of normal equations than that given above for the GLS-estimator of the parameter of the implicit EV model and that the derivation of the GLS-estimators for the explicit EV model requires certain modifications of that given for the GLS-estimators for the implicit EV model.

Thus in the explicit EV statistical model we are given an explicit linear structural relationship,

$$r(i) = \alpha_1 + \alpha_2 r_2(i) + \dots + \alpha_{p-1} r_{p-1}(i) ,$$

and we observe

$$X_2(i) = r_2(i) + \delta_2(i) ,$$

$$X_3(i) = r_3(i) + \delta_3(i) ,$$

$$\begin{array}{ccc} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{array}$$

$$X_{p-1}(i) = r_{p-1}(i) + \delta_{p-1}(i) ,$$

and

$$X(i) = r(i) + \delta(i) ,$$

where, for $1 \leq i \leq n$, the "errors"

$$\delta_*(i) = (\delta_2(i), \delta_3(i), \dots, \delta_{p-1}(i), \delta(i))$$

are assumed to have known covariance matrices given by

$$\Sigma_{ij} = E(\delta_*^T(i) \delta_*(j))$$

for $1 \leq i, j \leq n$.

We note that this explicit structural relationship can be written as an implicit one in the form

$$\alpha_1^* r_1^*(i) + \alpha_2^* r_2^*(i) + \dots + \alpha_{p-1}^* r_{p-1}^*(i) + \alpha_p^* r_p^*(i) = 0$$

where

$$\alpha_1^* = \alpha_1, \alpha_2^* = \alpha_2, \dots, \alpha_{p-1}^* = \alpha_{p-1}, \alpha_p^* = -1$$

and

$$r_1^*(i) = 1, r_2^*(i) = r_2(i), \dots, r_{p-1}^*(i) = r_{p-1}(i), r_p^*(i) = r(i) .$$

Let us write

$$\begin{aligned} r_*(i) &= (r_1^*(i), r_2^*(i), \dots, r_{p-1}^*(i), r_p^*(i)), \\ &= (1, r_2(i), \dots, r_{p-1}(i), r(i)) ; \\ \xi_*(i) &= (\xi_1^*(i), \xi_2^*(i), \dots, \xi_{p-1}^*(i), \xi_p^*(i)) \\ &= (0, \delta_2(i), \dots, \delta_{p-1}(i), \delta(i)) ; \\ X_*(i) &= (X_1^*(i), X_2^*(i), \dots, X_{p-1}^*(i), X_p^*(i)), \\ &= (1, X_2(i), \dots, X_{p-1}(i), X(i)) \end{aligned}$$

and

$$\begin{aligned} \alpha_* &= (\alpha_1^*, \alpha_2^*, \dots, \alpha_{p-1}^*, \alpha_p^*) \\ &= (\alpha_1, \alpha_2, \dots, \alpha_{p-1}, -1) ; \end{aligned}$$

then observing this notation, the EV statistical model in which the linear structural relationship is given explicitly can be written in vector form as

$$\alpha_* r_*(i) = 0$$

where, for $1 \leq i \leq n$, we observe

$$X_*(i) = r_*(i) + \xi_*(i)$$

and the covariance matrices of the "errors" $\xi_*(i)$ and $\xi_*(j)$ for $1 \leq i, j \leq n$ are given by

$$V_{ij} = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{ij} \end{bmatrix} = E(\xi_*^T(i) \xi_*(j)) \quad .$$

Thus we have reduced the explicit EV statistical model to an implicit EV statistical model in which certain prior information is known about the coefficients and variables of the implicit structural relationship and in which the covariance matrices of the "errors" have the special form of bordered matrices. In terms of this implicit EV model, which is equivalent to the given explicit EV model, we seek necessary conditions for an extreme point of the functional

$$Q(\alpha_*) = \sum_{r=1}^n \sum_{j=1}^n \alpha_* X_*^T(i) w_{ij}(\alpha_*) X_*(j) \alpha_*^T = d(\alpha_*) W(\alpha_*) d^T(\alpha_*)$$

where

$$W(\alpha_*) = [w_{ij}(\alpha_*)] = [\alpha_* V_{ij} \alpha_*^T]^{-1} = \Sigma_{d(\alpha_*)}^{-1}$$

and

$$d_i(\alpha_*) = \alpha_* X_*^T(i) \quad \text{for} \quad 1 \leq i \leq n$$

with

$$d(\alpha_*) = (d_1(\alpha_*), d_2(\alpha_*), \dots, d_n(\alpha_*)) \quad .$$

Now if we were to ignore the fact that α_p^* is a known constant and

treat it as an unknown parameter, we would be faced with the problem of finding necessary conditions for an extreme point of

$$Q(\tilde{\alpha}_*) = \sum_{i=1}^n \sum_{j=1}^n \tilde{\alpha}_* X_{*i}^T w_{ij}(\tilde{\alpha}_*) X_{*j} \tilde{\alpha}_*^T = d(\tilde{\alpha}_*) W(\tilde{\alpha}_*) d^T(\tilde{\alpha}_*)$$

where

$$W(\tilde{\alpha}_*) = [w_{ij}(\tilde{\alpha}_*)] = [\tilde{\alpha}_* V_{ij} \tilde{\alpha}_*^T]^{-1} = \Sigma_{d(\tilde{\alpha}_*)}^{-1}$$

and the tilda-notation

$$\tilde{\alpha}_* = (\tilde{\alpha}_1^*, \tilde{\alpha}_2^*, \dots, \tilde{\alpha}_{p-1}^*, \tilde{\alpha}_p^*)$$

denotes the vector

$$\alpha_* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_{p-1}^*, \alpha_p^*)$$

with α_p^* considered as an unknown parameter and hence as a variable along with $\alpha_1^*, \alpha_2^*, \dots$, and α_{p-1}^* .

But we have already obtained the solution to this problem in our derivation for the implicit EV model; and, therefore, the solution can be written down immediately from our previous work as

$$\left(\frac{\partial Q(\tilde{\alpha}_*)}{\partial \tilde{\alpha}_*} \right)^T = 2 A_{\tilde{\alpha}_*} \tilde{\alpha}_*^T = 0$$

where

$$A_{\tilde{\alpha}_*} = \sum_{r=1}^n \sum_{s=1}^n \left(w_{rs}(\tilde{\alpha}_*) \left(\frac{X_{*r}^T(r) X_{*s}(r) + X_{*s}^T(r) X_{*r}(r)}{2} \right) - \mu_r(\tilde{\alpha}_*) V_{rs} \mu_s(\tilde{\alpha}_*) \right) .$$

and

$$\mu(\tilde{\alpha}_*) = (\mu_1(\tilde{\alpha}_*), \mu_2(\tilde{\alpha}_*), \dots, \mu_n(\tilde{\alpha}_*)) = \tilde{\alpha}_* X_* W(\tilde{\alpha}_*)$$

and

$$X_* = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_2(1) & X_2(2) & \dots & X_2(n) \\ X_3(1) & X_3(2) & \dots & X_3(n) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ X_p(1) & X_p(2) & \dots & X_p(n) \end{bmatrix} = (X_*^T(1), X_*^T(2), \dots, X_*^T(n)) .$$

Actually, however, since α_p^* is in fact a known constant and not a variable, we really seek the Dwyer-Macphail matrix derivative of $Q(\alpha_*)$ with respect to the first $(p-1)$ -components of the vector

$$\alpha_* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_{p-1}^*, -1) ;$$

in other words, using a slight perversion of notation, we want

$$\frac{\partial Q(\alpha_*)}{\partial \alpha_*} = \left(\frac{\partial Q(\alpha_*)}{\partial \alpha_1^*}, \frac{\partial Q(\alpha_*)}{\partial \alpha_2^*}, \dots, \frac{\partial Q(\alpha_*)}{\partial \alpha_{p-1}^*} \right) .$$

But clearly, for $1 \leq k \leq p-1$,

$$\frac{\partial Q(\alpha_*)}{\partial \langle \alpha_* \rangle_k} = \left. \frac{\partial Q(\tilde{\alpha}_*)}{\partial \langle \tilde{\alpha}_* \rangle_k} \right]_{\tilde{\alpha}_p^* = -1} ;$$

and, therefore,

$$\frac{\partial Q(\alpha_*)}{\partial \alpha_*} = \left(\frac{\partial Q(\tilde{\alpha}_*)}{\partial \tilde{\alpha}_*} \right)_{\alpha_p^*=-1}^{\dagger}$$

where $\left(\frac{\partial Q(\tilde{\alpha}_*)}{\partial \tilde{\alpha}_*} \right)_{\alpha_p^*=-1}^{\dagger}$ denotes the vector $\frac{\partial Q(\tilde{\alpha}_*)}{\partial \tilde{\alpha}_*} = 2\tilde{\alpha}_* A \tilde{\alpha}_*$ with its p-th component deleted and with $\tilde{\alpha}_p^*$ replaced by the constant $\alpha_p^*=-1$.

Thus we have established the following

Theorem 4.3.2. Suppose the stochastic structure (f,S) is the EV statistical model where f is given explicitly by the linear structural relationship

$$r(i) = \alpha_1 + \alpha_2 r_2(i) + \dots + \alpha_{p-1} r_{p-1}(i)$$

and where only the random variables $X(i) = r(i) + \delta(i)$ and

$X_k(i) = r_k(i) + \delta_k(i)$ for $1 \leq k \leq p-1$ of the derived structure $D[S]$ of S are observable with the "errors" having zero means. Given a sample of n-observations, let $d(\alpha_*)$ be the vector whose i-th component is

$$d_i(\alpha_*) = \alpha_* X_*^T(i) \text{ where } \alpha_* = (\alpha_1, \alpha_2, \dots, \alpha_{p-1}, -1) \text{ and, for } 1 \leq i \leq n,$$

$X_*(i) = (1, X_2(i), X_3(i), \dots, X_{p-1}(i), X_p(i))$. Suppose that

$$W(\alpha_*) = [w_{ij}(\alpha_*)] = [\alpha_* V_{ij} \alpha_*^T]^{-1} = \Sigma_{d(\alpha_*)}^{-1}$$

exists for every α_* having at least two non-zero components where, for $1 \leq i, j \leq n$, the matrix V_{ij} is obtained by bordering the covariance matrix $\Sigma_{ij} = E(\delta_*^T(i) \delta_*(j))$ of the errors $\delta_*(i) = (\delta_2(i), \delta_3(i), \dots, \delta_{p-1}(i), \delta(i))$ by zeros. Then any non-trivial generalized least squares esti-

mator $(\alpha_1, \alpha_2, \dots, \alpha_{p-1})$ of $(\alpha_1, \alpha_2, \dots, \alpha_{p-1})$ must be a solution of the
non-linear equation system,

$$A_{\alpha_*}^{\dagger} \alpha_*^T = 0$$

where

$$A_{\alpha_*}^{\dagger} = \sum_{r=1}^n \sum_{s=1}^n \left(w_{rs}(\alpha_*) \left(\frac{X_*^T(r) X_*(s) + X_*^T(s) X_*(r)}{2} \right)^{\dagger} - \mu_r(\alpha_*) v_{rs}^{\dagger} \mu_s(\alpha_*) \right).$$

Here the dagger notation $A^{\dagger}$ denotes the matrix A with its last row de-
leted; and $\mu_k(\alpha_*) = \alpha_* X_* W(\alpha_*) e_k^T$ where X_* is the matrix whose i-th column
is the vector $X_*^T(i)$ for $1 \leq i \leq n$ and e_k is the n-dimensional row vector
having one in its k-th place and zero elsewhere.

Later, in the next chapter, we return to our GLS-solution of the estimation problem for the general implicit or explicit EV statistical models and find that solutions developed by other authors who made special assumptions about the probability distributions of the errors of the EV model (for example, the celebrated Gauss-Markoff-Aitken least squares estimators) can be obtained as special cases of our general results. At this point, however, it is convenient to look at the EVE statistical model and present a GLS-solution to the estimation problem for this stochastic structure using the methods of Dwyer-Macphail matrix differentiation.

4.4. GENERALIZED LEAST SQUARES ESTIMATORS FOR EVE MODEL

We now turn our attention to the problem of obtaining generalized

least squares estimators of the parameters of the EVE statistical model. First, however, we find it convenient as with the EV model to write the EVE model in vector notation. Keeping the notation introduced for the EV model, while introducing the following new notation;

$$\beta = (\beta_1, \beta_2, \dots, \beta_q),$$

$$S(i) = (S_1(i), S_2(i), \dots, S_q(i)) ,$$

$$\eta(i) = (\eta_1(i), \eta_2(i), \dots, \eta_q(i)) ,$$

and

$$Y(i) = (Y_1(i), Y_2(i), \dots, Y_q(i)) = S(i) + \eta(i),$$

we may write the EVE statistical model simply as

$$\alpha \cdot r^T(i) + \beta \cdot S^T(i) = \xi(i)$$

where, for $1 \leq i \leq n$, we observe

$$X(i) = r(i) + \delta(i) \quad \text{and} \quad Y(i) = S(i) + \eta(i) .$$

Moreover, writing $\sigma_{ij}^{(\xi)}$ instead of $\sigma_{\xi}(i, j)$, $\Sigma_{ij}^{(\eta)}$ instead of $\Sigma_{\eta}(i, j)$, $\Gamma_{ij}^{(\xi, \eta)}$ instead of $\Sigma_{(\xi, \eta)}(i, j)$, etc., the covariance matrix of the disturbance vectors $\gamma(i) = (\xi(i), \delta(i), \eta(i))$ and $\gamma(j) = (\xi(j), \delta(j), \eta(j))$ for $1 \leq i, j \leq n$ can be written as

$$\Sigma_{ij}^{(\gamma)} = \begin{bmatrix} \sigma_{ij}^{(\xi)} & \Gamma_{ij}^{(\xi, \delta)} & \Gamma_{ij}^{(\xi, \eta)} \\ \Gamma_{ij}^{(\delta, \xi)} & \Sigma_{ij}^{(\delta)} & \Sigma_{ij}^{(\delta, \eta)} \\ \Gamma_{ij}^{(\eta, \xi)} & \Sigma_{ij}^{(\eta, \delta)} & \Sigma_{ij}^{(\eta)} \end{bmatrix} = E(\gamma^T(i) \gamma(j)) .$$

We note that the covariance of $\xi(i)$ and $\xi(j)$ is related to the covariance matrix of the random vectors $S(i)$ and $S(j)$; in fact, since $E\xi(i) = 0$,

$$\alpha \cdot r^T(i) + \beta \cdot s^T(i) = E\xi(i) = 0 ,$$

where $s(i) = ES(i)$; and, therefore,

$$\beta \cdot s^T(i) = -\alpha \cdot r^T(i) .$$

Consequently,

$$\xi(i) = \beta(S^T(i) - s^T(i)) ,$$

and so

$$\sigma_{ij}^{(\xi)} = \beta \Sigma_{ij}^{(S)} \beta^T$$

or

$$\Sigma^{(\xi)} = [\sigma_{ij}^{(\xi)}] = [\beta \Sigma_{ij}^{(S)} \beta^T] .$$

In addition, upon multiplying both sides of the structural relationship of the EVE statistical model,

$$\alpha \cdot r^T(i) + \beta \cdot s^T(i) = \xi(i) ,$$

by the scalars $\delta(j)\alpha^T$ and $\eta(j)\beta^T$ and then taking expectations, we obtain two other important relationships among the variances and covariances of the random (vectors) variables of the EVE model, namely

$$\beta \sum_{ij}^{(S,\delta)} \alpha^T = \Gamma_{ij}^{(\xi,\delta)} \alpha^T$$

and

$$\beta \sum_{ij}^{(S,\eta)} \beta^T = \Gamma_{ij}^{(\xi,\eta)} \beta^T$$

for $1 \leq i, j \leq n$.

Having disposed of these introductory remarks we now proceed to set up the generalized least squares estimation criterion for the EVE statistical model. In this connection we introduce the "deviation,"

$$d_i^*(\theta) = \alpha \cdot X^T(i) + \beta \cdot Y^T(i) = \theta \cdot Z^T(i) ,$$

where $\theta = (\alpha, \beta)$, which corresponds to the i -th observation

$Z(i) = (X(i), Y(i))$ for $1 \leq i \leq n$. As in the case of the EV model we then consider the vector "deviation,"

$$d_*(\theta) = (d_1^*(\theta), d_2^*(\theta), \dots, d_n^*(\theta)) ,$$

formed from the n -scalar deviations $d_i^*(\theta)$. We note, moreover, that

$$Ed_*(\theta) = 0$$

since

$$\begin{aligned}
E(d_*(\theta)) &= \alpha EX^T(i) + \beta EY^T(i) \\
&= \alpha \cdot r^T(i) + \beta \cdot s^T(i) \\
&= E\xi(i) \\
&= 0 \quad .
\end{aligned}$$

In addition, the covariance matrix of the random vector $d_*(\theta)$ is given by

$$\Sigma_{d_*(\theta)} = [\sigma_{ij}^*(\theta)] = E(d_*^T(\theta)d_*(\theta))$$

where

$$\sigma_{ij}^*(\theta) = \theta \sum_{ij}^{(Z)} \theta^T$$

and

$$\sum_{ij}^{(Z)} = E((Z^T(i) - EZ^T(i))(Z(j) - EZ(j)))$$

for $1 \leq i, j \leq n$.

Now, writing $W_*(\theta) = \Sigma_{d_*(\theta)}^{-1}$ and assuming that $W_*(\theta)$ exists for every relevant θ in a domain Θ , the generalized least squares estimation procedure applied to the EVE statistical model leads us to the functional

$$Q_*(\theta) = d_*(\theta)W_*(\theta)d_*^T(\theta)$$

which is to be minimized with respect to the vector variable θ . Consequently, the GLS-criterion for the EVE model appears the same as the one for the EV model; and, therefore, at first glance, we might be led to

conclude that the normal equations for the generalized least squares estimators of the unknown parameters of the EVE model can be written down at once by making the proper identification with the normal equations for the GLS-estimators of the parameters of the EV model that were given in Section 4.3. This analysis is valid, of course, provided the covariance matrices $\Sigma_{ij}^{(S)}$, $\Sigma_{ij}^{(S,\delta)}$ and $\Sigma_{ij}^{(S,\eta)}$ are known since the matrix $\Sigma_{ij}^{(Z)}$ appearing in the expression $\sigma_{ij}^*(\theta) = \theta \Sigma_{ij}^{(Z)} \theta^T$ is given by

$$\Sigma_{ij}^{(Z)} = \begin{bmatrix} EX^T(i)X(j) - r^T(i)r(j) & EX^T(i)Y(j) - r^T(i)s(j) \\ EY^T(i)X(j) - s^T(i)r(j) & EY^T(i)Y(j) - s^T(i)s(j) \end{bmatrix}$$

where

$$EX^T(i)X(j) = r^T(i)r(j) + ES^T(i)\delta(j) ,$$

$$EX^T(i)Y(j) = r^T(i)ES(j) + ES^T(i)S(j) + ES^T(i)\eta(j) ,$$

$$EY^T(i)X(j) = ES^T(i)r(j) + ES^T(i)\delta(j) + E(\eta^T(i)\delta(j))$$

and

$$EY^T(i)Y(j) = ES^T(i)S(j) + ES^T(i)\delta(j) + E\eta^T(i)S(j) + E\eta^T(i)\eta(j) .$$

On the other hand, if the covariance matrices $\Sigma_{ij}^{(S)}$, $\Sigma_{ij}^{(S,\delta)}$ and $\Sigma_{ij}^{(S,\eta)}$ are not known, but rather only the covariance matrices of the equation and variable errors $\gamma(i) = (\xi(i), \delta(i), \eta(i))$ are known, since

$$\sigma_{ij}^{(\xi)} = \beta \Sigma_{ij}^{(S)} \beta^T ,$$

$$\beta \Sigma_{ij}^{(S, \delta)} \alpha^T = \Gamma_{ij}^{(\xi, \delta)} \alpha^T ,$$

and

$$\beta \Sigma_{ij}^{(S, \eta)} \beta^T = \Gamma_{ij}^{(\xi, \eta)} \beta^T ,$$

and α and β are unknown, the covariance matrices $\Sigma_{ij}^{(S)}$, $\Sigma_{ij}^{(S, \delta)}$ and $\Sigma_{ij}^{(S, \eta)}$ cannot be determined from knowledge of the covariance matrices of the errors $\gamma(i)$. Thus, under these circumstances, $\sigma_{ij}^*(\theta)$ cannot be expressed in the particular form $\sigma_{ij}^*(\theta) = \theta \Sigma_{ij}^{(Z)} \theta^T$ where $\Sigma_{ij}^{(Z)}$ is a known matrix; and for this reason, the derivation of the normal equations for the GLS-estimators of the unknown parameters of the EVE model departs from that given for the normal equations of the GLS-estimators of the EV model.

When only the covariance matrices of the equation and variable errors of the EVE statistical model are known, since

$$\begin{aligned} d_i^*(\theta) &= \alpha X^T(i) + \beta Y^T(i) \\ &= \alpha r^T(i) + \beta S^T(i) + \alpha \delta^T(i) + \beta \eta^T(i) \\ &= \xi(i) + \alpha \delta^T(i) + \beta \eta^T(i) , \end{aligned}$$

we find that

$$\sigma_{ij}^*(\theta) = \sigma_{ij}^{(\xi)} + \Gamma_{ij}^* \theta^T + \theta \Sigma_{ij}^* \theta^T = E(d_i^*(\theta) d_j^*(\theta))$$

where

$$\Gamma_{ij}^* = \Gamma_{ij} + \Gamma_{ji} ,$$

with

$$\Gamma_{ij} = E(\xi(i)(\delta(i), \eta(i))) = (\Gamma_{ij}^{(\xi, \delta)}, \Gamma_{ij}^{(\xi, \eta)})$$

and

$$\Sigma_{ij}^* = \begin{bmatrix} \Sigma_{ij}^{(\delta)} & \Sigma_{ij}^{(\delta, \eta)} \\ \Sigma_{ij}^{(\eta, \delta)} & \Sigma_{ij}^{(\eta)} \end{bmatrix} = E((\delta(i), \eta(i))^T (\delta(j), \eta(j))) .$$

Using this expression for $\sigma_{ij}^*(\theta)$ we now proceed to give a derivation of the normal equations for the GLS-estimators of the unknown (vector) parameter $\theta = (\alpha, \beta)$ of the EVE statistical model. Adopting the generalized least squares estimation procedure, therefore, the estimation problem in the EVE model may be stated as follows:

Using the Dwyer-Macphail matrix derivative calculus, obtain necessary conditions for a critical point of the functional

$$Q_*(\theta) = \sum_{i=1}^n \sum_{j=1}^n d_i^*(\theta) w_{ij}^*(\theta) d_j^*(\theta) = d_*(\theta) W_*(\theta) d_*^T(\theta)$$

where $d_*(\theta)$ is an n -dimensional row vector whose i -th component $d_i^*(\theta)$ for $1 \leq i \leq n$ is given by

$$d_i^*(\theta) = \alpha \cdot X^T(i) + \beta \cdot Y^T(i) = \theta \cdot Z^T(i) ,$$

$\theta = (\alpha, \beta)$ and $Z(i) = (X(i), Y(i))$, and where

$$W_*(\theta) = [w_{ij}^*(\theta)] = [\sigma_{ij}^*(\theta)]^{-1} = \Sigma_{d*}^{-1}(\theta)$$

with

$$\sigma_{ij}^*(\theta) = \sigma_{ij}^{(\xi)} + \Gamma_{ij}^* \theta^T + \theta \Sigma_{ij}^* \theta^T$$

is assumed to exist for every θ different from 0.

To carry out this derivation we write the functional $Q_*(\theta)$ in the form

$$Q_*(\theta) = \sum_{i=1}^n \sum_{j=1}^n f_{ij}^*(\theta) h_{ij}^*(\theta)$$

where

$$f_{ij}^*(\theta) = w_{ij}^*(\theta) \quad \text{and} \quad h_{ij}^*(\theta) = \theta Z^T(i) Z(j) \theta^T ;$$

then

$$\frac{\partial Q_*(\theta)}{\partial \theta} = \sum_{i=1}^n \sum_{j=1}^n \left(h_{ij}^*(\theta) \frac{\partial f_{ij}^*(\theta)}{\partial \theta} + f_{ij}^*(\theta) \frac{\partial h_{ij}^*(\theta)}{\partial \theta} \right) .$$

Clearly

$$\frac{\partial h_{ij}^*(\theta)}{\partial \theta} = \theta (Z^T(i) Z(j) + Z^T(j) Z(i)) ;$$

next, let

$$U_* = U_*(Y_*) = Y_*^{-1} \quad \text{and} \quad Y_* = V_*(\theta) = \Sigma_{d*}(\theta) ;$$

then

$$W_* = U_*(V_*(\theta)) = \Sigma_{d_*}^{-1}(\theta) = W_*(\theta) \quad .$$

Therefore, according to Theorem 2.4.4,

$$\frac{\partial w_{ij}^*(\theta)}{\partial \theta} = \sum_{r=1}^n \sum_{s=1}^n \left\langle \frac{\partial \langle U_* \rangle_{ij}}{\partial Y_*} \right\rangle_{rs} \left(\frac{\partial \langle Y_* \rangle_{rs}}{\partial \theta} \right) .$$

Since $\langle Y_* \rangle_{rs} = \sigma_{rs}^*(\theta) = \sigma_{rs}^{(\xi)} + \Gamma_{rs}^* \theta^T + \theta \Sigma_{rs}^* \theta^T$,

$$\frac{\partial \langle Y_* \rangle_{rs}}{\partial \langle \theta \rangle_k} = \Gamma_{rs}^* J_k^T + J_k \Sigma_{rs}^* \theta^T + \theta \Sigma_{rs}^* J_k^T ;$$

and, hence,

$$\frac{\partial \langle Y_* \rangle_{rs}}{\partial \theta} = \Gamma_{rs}^* + 2\theta \Sigma_{rs}^* .$$

Thus, using the identity

$$\left\langle \frac{\partial \langle U_* \rangle_{ij}}{\partial Y_*} \right\rangle_{rs} = - \left\langle (W_*(\theta))_i^T (W_*(\theta))_j \right\rangle_{rs}$$

which was also used in Section 4.3, we obtain for the Dwyer-Macphail matrix derivative of $w_{ij}^*(\theta)$ the expression

$$\begin{aligned} \frac{\partial w_{ij}^*(\theta)}{\partial \theta} = & - \sum_{r=1}^n \sum_{s=1}^n \left\langle (W_*(\theta))_i^T (W_*(\theta))_j \right\rangle_{rs} \Gamma_{rs}^* \\ & - 2\theta \left(\sum_{r=1}^n \sum_{s=1}^n \left\langle (W_*(\theta))_i^T (W_*(\theta))_j \right\rangle_{rs} \Sigma_{rs}^* \right) . \end{aligned}$$

Consequently,

$$\begin{aligned}
\frac{\partial \mathbf{q}_*(\theta)}{\partial \theta} = & 2\theta \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij}^*(\theta) \left(\frac{Z^T(i)Z(j) + Z^T(j)Z(i)}{2} \right) \right) \\
& - 2\theta \left(\sum_{i=1}^n \sum_{j=1}^n \theta Z^T(i)Z(j)\theta^T \left(\sum_{r=1}^n \sum_{s=1}^n < (W_*(\theta))_i^T (W_*(\theta))_j >_{rs} \Sigma_{rs}^* \right) \right) \\
& - \sum_{i=1}^n \sum_{j=1}^n \theta Z^T(i)Z(j)\theta^T \left(\sum_{r=1}^n \sum_{s=1}^n < (W_*(\theta))_i^T (W_*(\theta))_j >_{rs} \Gamma_{rs}^* \right) .
\end{aligned}$$

In turn, following the same derivation that was given in detail in Section 4.3, the following identities can be established:

$$\begin{aligned}
& \sum_{r=1}^n \sum_{s=1}^n \mu_r^*(\theta) \Sigma_{rs}^* \mu_s^*(\theta) \\
& = \sum_{i=1}^n \sum_{j=1}^n \theta Z^T(i)Z(j)\theta^T \left(\sum_{r=1}^n \sum_{s=1}^n < (W_*(\theta))_i^T (W_*(\theta))_j >_{rs} \Sigma_{rs}^* \right)
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{r=1}^n \sum_{s=1}^n \mu_r^*(\theta) \mu_s^*(\theta) \Gamma_{rs}^* \\
& = \sum_{i=1}^n \sum_{j=1}^n \theta Z^T(i)Z(j)\theta^T \left(\sum_{r=1}^n \sum_{s=1}^n < (W_*(\theta))_i^T (W_*(\theta))_j >_{rs} \Gamma_{rs}^* \right)
\end{aligned}$$

where

$$\mu_*(\theta) = (\mu_1^*(\theta), \mu_2^*(\theta), \dots, \mu_n^*(\theta)) = \theta Z W_*(\theta)$$

and

$$Z = [Z^T(1), Z^T(2), \dots, Z^T(n)] .$$

Using these identities, we see that the Dwyer-Macphail matrix derivative of the functional $Q_*(\theta)$ is given by the expression

$$\begin{aligned} \frac{\partial Q_*(\theta)}{\partial \theta} &= 2\theta \left(\sum_{r=1}^n \sum_{s=1}^n w_{rs}^*(\theta) \left(\frac{Z^T(r)Z(s) + Z^T(s)Z(r)}{2} \right) \right) \\ &\quad - 2\theta \left(\sum_{r=1}^n \sum_{s=1}^n \mu_r^*(\theta) \Sigma_{rs}^* \mu_s^*(\theta) \right) \\ &\quad - \sum_{r=1}^n \sum_{s=1}^n \mu_r^*(\theta) \mu_s^*(\theta) \Gamma_{rs}^* . \end{aligned}$$

Having derived this expression for $\partial Q_*(\theta)/\partial \theta$ the following theorem can be quickly established and its proof is omitted.

Theorem 4.4.1. Suppose the stochastic structure (f_*, S_*) is the EVE statistical model where f_* is given by the linear structural relationship $\alpha \cdot r^T(i) + \beta \cdot S^T(i) = \xi(i)$ and where only the random (vector) variables $X(i) = r(i) + \delta(i)$ and $Y(i) = S(i) + \eta(i)$ of the derived structure $D[S_*]$ of S_* are observable, while the equation "errors" $\xi(i)$ and the variable "errors" $\delta(i)$ and $\eta(i)$ are distributed about zero means. Given a sample of n-observations, let $d_*(\theta)$ be the vector whose i-th component is $d_i^*(\theta) = \alpha \cdot X^T(i) + \beta \cdot Y^T(i) = \theta \cdot Z^T(i)$ where $\theta = (\alpha, \beta)$ and $Z(i) = (X(i), Y(i))$ for $1 \leq i \leq n$. Suppose

$$W_*(\theta) = [w_{ij}^*(\theta)] = [\sigma_{ij}^*(\theta)]^{-1} = \Sigma_{d_*(\theta)}^{-1}$$

with

$$\sigma_{ij}^*(\theta) = \sigma_{ij}^{(\xi)} + \Gamma_{ij}^* \theta^T + \theta \Sigma_{ij}^* \theta^T$$

exists for every non-zero θ where the covariances $\sigma_{ij}^{(\xi)}$, $\Gamma_{ij}^* =$
 $E(\xi(i)(\delta(j), \eta(j))) + E(\xi(j)(\delta(i), \eta(i))) = \Gamma_{ij} + \Gamma_{ji}$ and $\Sigma_{ij}^* =$
 $E((\delta(i), \eta(i))^T (\delta(j), \eta(j)))$ of the equation and variable errors are known.
Then any non-trivial generalized least squares estimator $\hat{\theta} = (\hat{\beta}, \hat{\alpha})$ of
 $\theta = (\alpha, \beta)$ must be a solution of the non-linear equation system,

$$A_{\theta}^{*} \theta^T = \sum_{r=1}^n \sum_{s=1}^n \mu_r^*(\theta) \mu_s^*(\theta) \left(\frac{\Gamma_{rs} + \Gamma_{sr}}{2} \right)^T$$

where

$$A_{\theta}^{*} = \sum_{r=1}^n \sum_{s=1}^n \left(w_{ij}^*(\theta) \left(\frac{Z^T(r) Z(s) + Z^T(s) Z(r)}{2} \right) - \mu_r^*(\theta) \Sigma_{rs}^* \mu_s^*(\theta) \right),$$

and

$$\mu_*(\theta) = (\mu_1^*(\theta), \mu_2^*(\theta), \dots, \mu_n^*(\theta)) = \theta Z W_*(\theta),$$

and Z is the matrix whose i -th column is the vector of observations
 $Z^T(i)$ for $1 \leq i \leq n$.

The proof of this theorem is the same as the proof given for Theorem 4.3.1, but using the Dwyer-Macphail matrix derivative $\partial Q_*(\theta)/\partial \theta$ of the GLS-functional $Q_*(\theta)$ associated with the EVE model instead of the Dwyer-Macphail matrix derivative $\partial Q(\alpha)/\partial \alpha$ of the GLS-functional $Q(\alpha)$ based on the EV model. We readily see, moreover, when $\sigma^2 \xi(i) = 0$, $\sigma^2 S_k(i) = 0$ and $\sigma^2 \eta_k(j) = 0$ for $1 \leq i \leq n$ and $1 \leq k \leq q$ or, equivalently, when $\xi(i) = 0$, $S_k(j) = 0$ and $\eta_k(j) = 0$ with probability one, that Theorem 4.4.1 reduces to Theorem 4.3.1 provided $s_k(j) = 0$.

By a direct application of the Dwyer-Macphail matrix derivative calculus, we have derived the systems of non-linear "normal" equations which define implicitly the generalized least squares estimators of the unknown parameters of a linear structural relationship when all of its variables are non-observable (i.e., for the error-in-variable statistical model) or when "shocks" or equation errors are present (i.e., for the error-in-variable and/or error-in-equation statistical model). These applications of the calculus of Dwyer-Macphail matrix derivatives have provided simple but elegant derivations of the normal equations representing the necessary conditions for a critical point of two complicated non-linear functionals of vector variables. These functionals can be interpreted as generalized least squares criterion based on systems of weights in which the weights, as well as the deviations of the observations from the underlying structural relationship, depend on the unknown parameters to be estimated.

4.5. REGULARITY CONDITIONS

We begin this section by studying regularity conditions involving linear combinations of the variable and equation errors of the EV and EVE statistical models that are equivalent to the assumptions made in Theorems 4.3.1, 4.3.2 and 4.4.1 that the matrices $W(\alpha) = \Sigma_d^{-1}(\alpha)$, $W(\alpha_*) = \Sigma_d^{-1}(\alpha_*)$ and $W_*(\theta) = \Sigma_d^{-1}(\theta)$ exist for every non-zero α , non-zero θ or for every α_* having at least one non-zero component besides α_p^* . These regularity conditions stated in terms of the variable and equation

errors of the EV and EVE models, although equivalent to the assumption that the matrices $\Sigma_d(\alpha)$, $\Sigma_d(\alpha_*)$, and $\Sigma_{d*}(\theta)$ are positive definite matrices, frame the question of the existence of $W(\alpha)$, $W(\alpha_*)$ and $W_*(\theta)$ in terms of restrictions on the random variables of the underlying stochastic structures S and S_* which generate the EV and EVE models and thus returns our point of view and emphasis to these more basic constructions.

We know that $\Sigma_d(\alpha) = [\sigma_{ij}(\alpha)]$ where $\sigma_{ij}(\alpha) = \alpha \sum_{1j} \alpha^T$ and that $\Sigma_{d*}(\theta) = [\sigma_{ij}^*(\theta)]$ where $\sigma_{ij}^*(\theta) = \theta \sum_{1j}^{(Z)} \theta^T$ or $\sigma_{ij}^*(\theta) = \sigma_{ij}^{(\xi)} + \Gamma_{ij}^* \theta^T + \theta \sum_{1j} \theta^T$, as covariance matrices, are real, symmetric matrices. Moreover, since for every vector $a = (a_1, a_2, \dots, a_n)$ and every $\alpha \neq 0$ or $\theta \neq 0$,

$$a \Sigma_d(\alpha) a^T = \sum_{i=1}^n \sum_{j=1}^n a_i \sigma_{ij}(\alpha) a_j = E \left(\sum_{i=1}^n a_i (d_i(\alpha) - E d_i(\alpha)) \right)^2 \geq 0$$

and

$$a \Sigma_{d*}(\theta) a^T = \sum_{i=1}^n \sum_{j=1}^n a_i \sigma_{ij}^*(\theta) a_j = E \left(\sum_{i=1}^n a_i (d_i^*(\theta) - E d_i^*(\theta)) \right)^2 \geq 0,$$

we know that $\Sigma_d(\alpha)$ and $\Sigma_{d*}(\theta)$ are either positive semi-definite or positive definite matrices for every α and θ . As the following theorems indicate, these matrices are in fact positive definite uniformly in α or θ provided, with probability one, no linear combinations of the errors of the EV and EVE statistical models are present.

Theorem 4.5.1. If at least one of the covariance matrices

$\Sigma_{ii} = E \delta^T(i) \delta(j)$ of the errors of the implicit EV statistical model is positive definite for $1 \leq i \leq n$, then the matrix $\Sigma_d(\alpha)$ is positive defi-

nite for every non-zero α if and only if $|\rho_j(\lambda, \mu)| < 1$ for $1 \leq j \leq n$
and every non-zero p-dimensional vector λ and every non-zero n-dimen-
sional vector μ . Here $\rho_j(\lambda, \mu)$ denotes the correlation coefficient of
the random variables $\lambda \delta^T(j)$ and $\lambda \delta^T(\mu; j)$ where $\delta(\mu; j) = \sum_{k \neq j} \mu_k \delta(k)$.

Proof. According to a theorem in Cramér [24], we know that $\Sigma_d(\bar{\alpha})$ is of rank $r < n$ if and only if there exists $m = n - r > 0$ linearly independent vectors

$$a_k = (a_1^{(k)}, a_2^{(k)}, \dots, a_n^{(k)})$$

such that, with probability one,

$$\sum_{i=1}^n a_i^{(k)} (d_i(\bar{\alpha}) - E d_i(\bar{\alpha})) = \sum_{i=1}^n a_i^{(k)} \bar{\alpha} \delta^T(i) = 0,$$

for $1 \leq k \leq m$. Now, at least one of the vectors a_k cannot be of the form $a_k = a e_k$ where a is a non-zero scalar and e_k is the n -dimensional vector having one in its k -th place and zero elsewhere; for if each a_k were of this form it would imply that $\bar{\alpha} \delta^T(k) = 0$ and, therefore, for $1 \leq k \leq n$, $\bar{\alpha} \sum_{kk} \bar{\alpha}^T = 0$ for a non-zero vector $\bar{\alpha}$ which contradicts the assumption that some $\sum_{kk}$ is a positive definite matrix. Consequently, if $\Sigma_d(\bar{\alpha})$ were of rank $r < n$, since at least one a_k has more than one non-zero component, it would be possible to express some $\bar{\alpha} \delta^T(j)$ as a linear combination of other $\bar{\alpha} \delta^T(k)$'s say

$$\bar{\alpha} \delta^T(j_0) = \sum_{k \neq j_0} \mu_k \bar{\alpha} \delta^T(k);$$

and, therefore, taking $\lambda = \bar{\alpha}$ and $\mu = \bar{\mu}$, we would have $|\rho_{j_0}(\bar{\alpha}, \bar{\mu})| = 1$ which is a contradiction. Hence if $|\rho_j(\lambda, \mu)| < 1$ for $1 \leq j \leq n$ and every non-zero λ and μ , then $\Sigma_d(\alpha)$ must have rank n for every non-zero α and is a positive definite matrix. Conversely, suppose $\Sigma_d(\alpha)$ is of rank n uniformly in $\alpha \neq 0$; then for every non-zero $\lambda = \alpha$ and μ ,

$$\sum_{k=1}^n \mu_k \lambda \delta^T(k) \neq 0,$$

and hence $|\rho_j(\lambda, \mu)| < 1$ for $1 \leq j \leq n$. Thus the proof of Theorem 4.5.1 is completed.

A corresponding result for the explicit EV model can be obtained by making certain obvious modifications in the statement of Theorem 4.5.1. For the sake of completeness, we restate Theorem 4.5.1 with these modifications as Theorem 4.5.2, but its proof being identical with that given for the implicit EV model is omitted.

Theorem 4.5.2. If at least one of the covariance matrices
 $\Sigma_{ii} = E\delta_*^T(i)\delta_*^T(j)$ of the errors $\delta_*(i)$ of the explicit EV statistical
model is positive definite for $1 \leq i \leq n$, then the matrix $\Sigma_d(\alpha_*)$ is
positive definite for every α_* having at least one non-zero component
besides α_p^* if and only if $|\rho_j(\lambda_*; \mu)| < 1$ for $1 \leq j \leq n$ and every non-
zero $(p-1)$ -dimensional vector λ_* and every non-zero n -dimensional vec-
tor μ . Here $\rho_j(\lambda_*; \mu)$ is the correlation coefficient of the random
variables $\lambda_*\delta_*^T(j)$ and $\lambda_*\delta_*^T(\mu; j)$ where $\delta_*(\mu; j) = \sum_{k \neq j} \mu_k \delta_*(k)$.

We now present a similar result for the EVE statistical model. Let us denote the variable errors $(\delta(i), \eta(i))$ for $1 \leq i \leq n$ of the EVE

statistical model by $\psi(i)$ and for any $(p+q)$ -dimensional vector let $\psi_\mu(i) = \mu\psi^T(i)$; finally, let $\psi_\mu = (\psi_\mu(1), \psi_\mu(2), \dots, \psi_\mu(n))$ and $\xi = (\xi(1), \xi(2), \dots, \xi(n))$ where the $\xi(i)$'s are the equation errors of the EVE model. Then we state

Theorem 4.5.3. The covariance matrix $\Sigma_{d_*}(\theta)$ of the EVE model deviations $d_i^*(\theta) = \alpha X^T(i) + \beta Y^T(i) = \theta Z^T(i)$ is positive definite for every non-zero θ if and only if $|\rho(\lambda_\xi, \lambda_{\psi_\mu})| < 1$ for every non-zero n -dimensional vector λ and every non-zero $(p+q)$ -dimensional vector μ . Here ρ denotes the correlation coefficient of the random variables λ_ξ and λ_{ψ_μ} .

Proof. Suppose $\Sigma_{d_*}(\theta)$ has rank $r < n$ for a particular non-zero $\theta = \bar{\theta}$. Then there exists a non-zero vector $a = (a_1, a_2, \dots, a_n)$ such that

$$\sum_{i=1}^n a_i (d_i^*(\bar{\theta}) - E d_i^*(\bar{\theta})) = \sum_{i=1}^n a_i (\xi(i) + \bar{\theta} \psi^T(i)) = 0.$$

Thus we find that

$$\sum_{i=1}^n a_i \xi(i) = - \sum_{i=1}^n a_i \bar{\theta} \psi^T(i)$$

and therefore $|\rho(a_\xi, a_{\psi \bar{\theta}})| = 1$ which is a contradiction. Hence whenever $|\rho(\lambda_\xi, \lambda_{\psi_\mu})| < 1$ for every non-zero λ and μ , then $\Sigma_{d_*}(\theta)$ must have rank n for every non-zero θ and thus is positive definite. Conversely, if $\Sigma_{d_*}(\theta)$ is positive definite for every non-zero θ , it is of rank n for every non-zero θ . Hence

$$\sum_{i=1}^n \lambda_i (\xi(i) + \mu \psi^T(i)) = \sum_{i=1}^n \lambda_i (d_i^*(\mu) - E d_i^*(\mu)) \neq 0$$

for any non-zero λ or non-zero μ and therefore $|\rho(\lambda_\xi, \lambda\psi_\mu)| < 1$ which completes the proof of Theorem 4.5.3.

Having related the existence of the matrices $W(\alpha)$, $W(\alpha_*)$ and $W_*(\theta)$ to regularity conditions involving restrictions on linear combinations of the variable and equation errors of the EV and EVE statistical models, we now examine one more point related to the question of what regularity conditions among the observations and covariances of the errors of the EV and EVE models are needed in the form of a sufficient condition that the GLS-functionals $Q(\alpha)$ or $Q_*(\theta)$ have a local minimum at a critical point and thus are truly GLS-estimators of the parameters α or θ . We search for such a regularity condition first under the assumption that the covariance matrices $\Sigma_d(\alpha)$ and $\Sigma_{d*}(\theta)$ are diagonal and then extend these results to non-diagonal covariance matrices. The proofs given below draw heavily upon the relationships that exist between the Dwyer-Macphail matrix derivatives and Frechet-Gateaux matrix differentials as developed in Chapter 3, and we use these results freely and generally without explicit reference to them.

Consider the EV model where we assume that $\Sigma_d(\alpha)$ is a diagonal matrix and therefore

$$W(\alpha) = \begin{bmatrix} \frac{1}{\sigma_1(\alpha)} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2(\alpha)} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & \frac{1}{\sigma_n(\alpha)} \end{bmatrix}$$

where $\sigma_i(\alpha) = \alpha \sum_{i=1}^n \alpha^T$ and $\sum_{i=1}^n = E(\delta^T(i)\delta(i))$ is the covariance matrix of the variable errors $\delta(i)$ of the EV model for $1 \leq i \leq n$.

Then our GLS-functional can be written as

$$Q(\alpha) = \sum_{i=1}^n \frac{(\alpha X^T(i))^2}{\sigma_i(\alpha)}.$$

The Frechet-Gateaux matrix differential of $Q(\alpha)$ is given by the expression

$$\delta Q(\alpha; k) = \sum_{i=1}^n \frac{2(\alpha X^T(i))(k X^T(i))}{\sigma_i(\alpha)} - \sum_{i=1}^n \left(\frac{\alpha X^T(i)}{\sigma_i(\alpha)} \right)^2 \delta \sigma_i(\alpha; k)$$

where $\delta \sigma_i(\alpha; k) = 2k \sum_{i=1}^n \alpha^T$. Let $\alpha_0 \neq 0$ be a point such that $\delta Q(\alpha_0; k) = 0$ identically in k ; in other words

$$2 \sum_{i=1}^n \frac{(\alpha_0 X^T(i))(k X^T(i))}{\sigma_i(\alpha_0)} = \sum_{i=1}^n \left(\frac{\alpha_0 X^T(i)}{\sigma_i(\alpha_0)} \right)^2 \delta \sigma_i(\alpha_0; k).$$

Since $\delta^2 Q(\alpha; k) = \delta_k \delta Q(\alpha; k)$, another direct computation leads to the following expression for the second order Frechet-Gateaux matrix differential of $Q(\alpha)$, namely

$$\begin{aligned} \delta^2 Q(\alpha; k) &= \sum_{i=1}^n \frac{2(k X^T(i))^2}{\sigma_i(\alpha)} + \sum_{i=1}^n \frac{2(\alpha X^T(i))^2}{\sigma_i^3(\alpha)} (\delta \sigma_i(\alpha; k))^2 \\ &\quad - \sum_{i=1}^n \frac{4(\alpha X^T(i))(k X^T(i))}{\sigma_i^2(\alpha)} \delta \sigma_i(\alpha; k) - \sum_{i=1}^n \left(\frac{\alpha X^T(i)}{\sigma_i(\alpha)} \right)^2 \delta^2 \sigma_i(\alpha; k) \end{aligned}$$

where $\delta^2 \sigma_i(\alpha; k) = 2k \sum_{i=1}^n k^T$. This expression for $\delta^2 Q(\alpha; k)$ can be rewritten in the form

$$\begin{aligned} \delta^2 Q(\alpha; k) &= 2kXW(\alpha)X^T k^T + 2\alpha XW(\alpha) [\delta\sigma_1(\alpha; k)]W(\alpha) [\delta\sigma_1(\alpha; k)]W(\alpha)X\alpha^T \\ &\quad - 4kXW(\alpha) [\delta\sigma_1(\alpha; k)]W(\alpha)X^T \alpha^T - 2\alpha XW(\alpha) [k \sum_{i=1}^n k^T]W(\alpha)X^T \alpha^T, \end{aligned}$$

where X denotes the **matrix** whose **columns** are the vectors $X^T(i)$ for $1 \leq i \leq n$.

Using properties of the Tr -operator together with the fact that $\delta\sigma_1(\alpha; k) = \left(\frac{\partial\sigma_1(\alpha)}{\partial\alpha} \right) k^T$ where $\frac{\partial\sigma_1(\alpha)}{\partial\alpha}$ is the Dwyer-Macphail matrix derivative of $\sigma_1(\alpha)$, the following identity can be established:

$$\begin{aligned} &\alpha XW(\alpha) [\delta\sigma_1(\alpha; k)]W(\alpha) [\delta\sigma_1(\alpha; k)]W(\alpha)X^T \alpha^T \\ &= \text{Tr} [\delta\sigma_1(\alpha; k)]W(\alpha) [\delta\sigma_1(\alpha; k)] [W(\alpha)X^T \alpha^T \alpha XW(\alpha)] \\ &= \text{Tr} \left[\left(\sum_{i=1}^n \delta\sigma_1(\alpha; k) E_{ii} \right) W(\alpha) \left(\sum_{j=1}^n \delta\sigma_j(\alpha; k) E_{jj} \right) \right] [W(\alpha)X^T \alpha^T \alpha XW(\alpha)] \\ &= \text{Tr} \left[\sum_{i=1}^n \sum_{j=1}^n \delta\sigma_1(\alpha; k) E_{ii} W(\alpha) E_{jj} \delta\sigma_j(\alpha; k) \right] [W(\alpha)X^T \alpha^T \alpha XW(\alpha)] \end{aligned}$$

where E_{ij} denotes the $n \times n$ -matrix having one in its (i, j) -th place and zeros elsewhere. Since $E_{ii} W(\alpha) E_{jj} = w_{ij}(\alpha) E_{ij}$ and $w_{ij}(\alpha) = 0$ when $i \neq j$, we find that

$$\alpha X W(\alpha) [\delta \sigma_i(\alpha; k)] W(\alpha) [\delta \sigma_i(\alpha; k)] W(\alpha) X^T \alpha^T$$

$$\begin{aligned}
 &= \text{Tr} \left[\left(\sum_{i=1}^n w_{ii}(\alpha) (\delta \sigma_i(\alpha; k))^2 \right) E_{ii} \right] [W(\alpha) X^T \alpha^T \alpha X W(\alpha)] \\
 &= \sum_{i=1}^n w_{ii}(\alpha) (\delta \sigma_i(\alpha; k))^2 \text{Tr}(\alpha X W(\alpha) E_{ii} W(\alpha) X^T \alpha^T) \\
 &= \sum_{i=1}^n w_{ii}(\alpha) \left(k \left(\frac{\partial \sigma_i}{\partial \alpha} \right)^T \left(\frac{\partial \sigma_i}{\partial \alpha} \right) k^T \right) (\alpha X W(\alpha) E_{ii} W(\alpha) X^T \alpha^T) \\
 &= k \left[\sum_{i=1}^n \frac{(\alpha X^T(i))^2}{\sigma_i^3(\alpha)} \left(\frac{\partial \sigma_i}{\partial \alpha} \right)^T \left(\frac{\partial \sigma_i}{\partial \alpha} \right) \right] k^T.
 \end{aligned}$$

By similar arguments the following identities can also be established:

$$k X W(\alpha) [\delta \sigma_i(\alpha; k)] W(\alpha) X^T \alpha^T = k \left[\sum_{i=1}^n \left(\frac{\alpha X^T(i)}{\sigma_i^2(\alpha)} \right) \left(\frac{\partial \sigma_i}{\partial \alpha} \right)^T X(i) \right] k^T ;$$

and

$$\alpha X W(\alpha) [k \sum_{i=1}^n k^T] W(\alpha) X^T \alpha^T = k \left[\sum_{i=1}^n \left(\frac{\alpha X^T(i)}{\sigma_i(\alpha)} \right)^2 \sum_{i=1}^n \right] k^T.$$

Combining these results we find that $\delta^2 Q(\alpha; k)$ can be written in the form

$$\delta^2 Q(\alpha; k) = 2k B_\alpha k^T$$

where

$$B_{\alpha} = \sum_{i=1}^n \left[\frac{X^T(i)X(i)}{\sigma_i(\alpha)} + \frac{(\alpha X^T(i))^2}{\sigma_i^3(\alpha)} \left(\frac{\partial \sigma_i}{\partial \alpha} \right)^T \left(\frac{\partial \sigma_i}{\partial \alpha} \right) - 2 \frac{\alpha X^T(i)}{\sigma_i^2(\alpha)} \left(\frac{\partial \sigma_i}{\partial \alpha} \right)^T X(i) - \left(\frac{\alpha X^T(i)}{\sigma_i(\alpha)} \right)^2 \Sigma_{ii} \right].$$

Recalling Theorem 3.5.5 and keeping in mind this expression for $\delta^2 Q(\alpha; k)$

we see that we have established the following result:

Theorem 4.5.4. If $\delta Q(\alpha_0; k) = 0$ identically in k (i.e., if α_0 is a critical point of $Q(\alpha) = \alpha XW(\alpha)X^T\alpha^T = d(\alpha)W(\alpha)d^T(\alpha)$ where $W(\alpha)$ is a diagonal matrix whose i -th diagonal element is $1/\sigma_i(\alpha)$ with $\sigma_i(\alpha) = \alpha \Sigma_{ii}\alpha^T$) then α_0 is actually a minimum point of $Q(\alpha)$ provided the matrix B_{α_0} is positive definite where

$$B_{\alpha_0} = \sum_{i=1}^n \left[\frac{X^T(i)X(i)}{\sigma_i(\alpha_0)} + \frac{(\alpha_0 X^T(i))^2}{\sigma_i^3(\alpha_0)} \left(\frac{\partial \sigma_i}{\partial \alpha_0} \right)^T \left(\frac{\partial \sigma_i}{\partial \alpha_0} \right) - 2 \frac{\alpha_0 X^T(i)}{\sigma_i^2(\alpha_0)} \left(\frac{\partial \sigma_i}{\partial \alpha_0} \right)^T X(i) - \left(\frac{\alpha_0 X^T(i)}{\sigma_i(\alpha_0)} \right)^2 \Sigma_{ii} \right]$$

and $\partial \sigma_i / \partial \alpha_0$ is the Dwyer-Macphail derivative of $\sigma_i(\alpha)$ evaluated at

$\alpha = \alpha_0$.

The arguments that led to this result depended only on the form of the functional $Q(\alpha) = \alpha XW(\alpha)X^T\alpha^T = d(\alpha)W(\alpha)d^T(\alpha)$ and the fact that $\delta \sigma_i(\alpha; k) = (\partial \sigma_i / \partial \alpha) k^T$ and that $\delta^2 \sigma_i(\alpha; k) = 2k \Sigma_{ii} k^T$. Thus they remain valid if in place of the GLS-functional $Q(\alpha)$ for the EV model we take the GLS-functional $Q_*(\theta) = \theta ZW_*(\theta)Z^T\theta^T = d_*(\theta)W_*(\theta)d_*^T(\theta)$ for the EVE

model while noting that $\delta\sigma_i^*(\theta;h) = (\partial\sigma_i^*/\partial\theta) h^T$ and $\delta^2\sigma_i^*(\theta;h) = 2h\sum_{ii}^* h^T$. Consequently we have the following theorem for the EVE statistical model,

Theorem 4.5.5. If $\delta Q_*(\theta_0;h) = 0$ identically in h (i.e., if θ_0 is a critical point of $Q_*(\theta) = \theta ZW(\theta)Z^T\theta^T = d_*(\theta)W_*(\theta)d_*^T(\theta)$ where $W_*(\theta)$ is a diagonal matrix whose i -th diagonal element is $1/\sigma_i^*(\theta)$ with $\sigma_i^*(\theta) = \sigma_{ii}^{(\xi)} + \Gamma_{ii}^*\theta^T + \theta\sum_{ii}^*\theta^T$) then θ_0 is actually a minimum point of $Q_*(\theta)$ provided the matrix $B_{\theta_0}^*$ is positive definite where

$$B_{\theta_0}^* = \sum_{i=1}^n \left[\frac{Z^T(i)Z(i)}{\sigma_i^*(\theta_0)} + \frac{(\theta_0 Z^T(i))^2}{(\sigma_i^*(\theta_0))^3} \left(\frac{\partial\sigma_i^*}{\partial\theta_0} \right)^T \left(\frac{\partial\sigma_i^*}{\partial\theta_0} \right) - 2 \frac{\theta_0 Z^T(i)}{(\sigma_i^*(\theta_0))^2} \left(\frac{\partial\sigma_i^*}{\partial\theta_0} \right)^T Z(i) - \left(\frac{\theta_0 Z^T(i)}{\sigma_i^*(\theta_0)} \right)^2 \sum_{ii}^* \right]$$

and $\partial\sigma_i^*/\partial\theta_0$ is the Dwyer-Macphail matrix derivative of $\sigma_i^*(\theta)$ evaluated at $\theta = \theta_0$.

We now go on to extend these two theorems to the case of non-diagonal covariances $\sum_d(\alpha)$ and $\sum_{d_*}(\theta)$ for the EV and EVE models. Thus consider the GLS-functional

$$Q(\alpha) = \alpha XW(\alpha)X^T\alpha^T = d(\alpha)W(\alpha)d^T(\alpha)$$

where now $W(\alpha)$ is no longer a diagonal matrix. The Frechet-Gateaux matrix differential of $Q(\alpha)$ is given by the expression

$$\delta Q(\alpha;k) = 2kXW(\alpha)X^T\alpha^T - \alpha XW(\alpha)\delta\sum(\alpha;k)W(\alpha)X^T\alpha^T$$

where $\delta \Sigma(\alpha; k)$ denotes the Frechet-Gateaux matrix differential of $\Sigma_d(\alpha)$.

Now $\alpha_0 \neq 0$ is a critical point of $Q(\alpha)$ provided $\delta Q(\alpha_0; k) = 0$ identically in k ; or, equivalently, provided for every $k \neq 0$,

$$2kXW(\alpha_0)X^T\alpha_0^T = \alpha_0XW(\alpha_0)\delta\Sigma(\alpha_0; k)W(\alpha_0)X^T\alpha_0^T.$$

Next, since $\delta^2 Q(\alpha; k) = \delta_k \delta Q(\alpha; k)$, we find that the second order Frechet-Gateaux differential of $Q(\alpha)$ is given by

$$\delta^2 Q(\alpha; k) = 2kXW(\alpha)X^T k^T + 4kX\delta W(\alpha; k)X^T\alpha^T + \alpha X\delta^2 W(\alpha, k)X^T\alpha^T$$

where

$$\delta W(\alpha; k) = -W(\alpha)\delta\Sigma(\alpha; k)W(\alpha)$$

and

$$\delta^2 W(\alpha; k) = 2W(\alpha)\delta\Sigma(\alpha; k)W(\alpha)\delta\Sigma(\alpha; k)W(\alpha) - W(\alpha)\delta^2\Sigma(\alpha; k)W(\alpha).$$

Combining these expressions we finally obtain

$$\begin{aligned}\delta^2 Q(\alpha; k) &= 2kXW(\alpha)X^T k^T - \alpha XW(\alpha)\delta^2\Sigma(\alpha; k)W(\alpha)X^T\alpha^T \\ &\quad + 2\alpha XW(\alpha)\delta\Sigma(\alpha; k)W(\alpha)\delta\Sigma(\alpha; k)W(\alpha)X^T\alpha^T \\ &\quad - 4kXW(\alpha)\delta\Sigma(\alpha; k)W(\alpha)X^T\alpha^T.\end{aligned}$$

Using the fact that

$$\delta\Sigma(\alpha; k) = [\delta\sigma_{1j}(\alpha; k)] \quad \text{where} \quad \delta\sigma_{1j}(\alpha; k) = \left(\frac{\partial\sigma_{1j}}{\partial\alpha}\right) k^T$$

or that

$$\delta^2\Sigma(\alpha; k) = [\delta^2\sigma_{1j}(\alpha; k)] \quad \text{where} \quad \delta^2\sigma_{1j}(\alpha; k) = 2k\Sigma_{1j}k^T,$$

together with properties of the Tr-operator, we can establish the following identities:

$$\alpha XW(\alpha) \delta^2 \sum(\alpha; k) W(\alpha) X^T \alpha^T = 2k \left[\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \Sigma_{ij} \right] k^T ;$$

$$\begin{aligned} & \alpha XW(\alpha) \delta \sum(\alpha; k) W(\alpha) \delta \sum(\alpha; k) W(\alpha) X^T \alpha^T \\ &= k \left[\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \sum_{r=1}^n \sum_{s=1}^n w_{rs}(\alpha) \left(\frac{\partial \sigma_{ri}}{\partial \alpha} \right)^T \left(\frac{\partial \sigma_{sj}}{\partial \alpha} \right) \right] k^T ; \end{aligned}$$

and

$$k XW(\alpha) \delta \sum(\alpha; k) W(\alpha) X^T \alpha^T = k \left[\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \sum_{s=1}^n w_{sj}(\alpha) \left(\frac{\partial \sigma_{ij}}{\partial \alpha} \right)^T X(s) \right] k^T .$$

Proofs for these identities follow; for example,

$$\begin{aligned} \alpha XW(\alpha) \delta^2 \sum(\alpha; k) W(\alpha) X^T \alpha^T &= \text{Tr}[\delta^2 \sigma_{ij}(\alpha; k)] [W(\alpha) X^T \alpha^T \alpha XW(\alpha)] \\ &= \text{Tr} \left[\sum_{i=1}^n \sum_{j=1}^n \delta^2 \sigma_{ij}(\alpha; k) E_{ij} \right] [W(\alpha) X^T \alpha^T \alpha XW(\alpha)] \\ &= \sum_{i=1}^n \sum_{j=1}^n \delta^2 \sigma_{ij}(\alpha; k) \text{Tr}(\alpha XW(\alpha) E_{ij}^T W(\alpha) X^T \alpha^T) \\ &= 2k \left[\sum_{i=1}^n \sum_{j=1}^n (\alpha XW(\alpha) E_{ji} W(\alpha) X^T \alpha^T) \Sigma_{ij} \right] k^T \\ &= 2k \left[\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \Sigma_{ij} \right] k^T \end{aligned}$$

where E_{ij} denotes the $n \times n$ -matrix having one in its (i,j) -th position and zeros elsewhere and $\mu(\alpha) = (\mu_1(\alpha), \mu_2(\alpha), \dots, \mu_n(\alpha)) = \alpha X W(\alpha)$.

Next,

$$\begin{aligned}
 & \alpha X W(\alpha) \delta \Sigma(\alpha; k) W(\alpha) \delta \Sigma(\alpha; k) W(\alpha) X^T \alpha^T \\
 &= \text{Tr}[\delta \Sigma(\alpha; k) W(\alpha) \delta \Sigma(\alpha; k)] [W(\alpha) X^T \alpha^T \alpha X W(\alpha)] \\
 &= \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \sum_{s=1}^n \text{Tr}[\delta \sigma_{ri}(\alpha; k) E_{ri} W(\alpha) E_{sj} \delta \sigma_{sj}(\alpha; k)] [W(\alpha) X^T \alpha^T \alpha X W(\alpha)] \\
 &= \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \sum_{s=1}^n \text{Tr}[\delta \sigma_{ri}(\alpha; k) w_{rs}(\alpha) \delta \sigma_{sj}(\alpha; k) E_{ij}] [W(\alpha) X^T \alpha^T \alpha X W(\alpha)] \\
 &= \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \sum_{s=1}^n \delta \sigma_{ri}(\alpha; k) w_{rs}(\alpha) \delta \sigma_{sj}(\alpha; k) \text{Tr}(\alpha X W(\alpha) E_{ij}^T W(\alpha) X^T \alpha^T) \\
 &= \sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \sum_{r=1}^n \sum_{s=1}^n \delta \sigma_{ri}(\alpha; k) w_{rs}(\alpha) \delta \sigma_{sj}(\alpha; k) \\
 &= k \left[\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \sum_{r=1}^n \sum_{s=1}^n w_{rs}(\alpha) \left(\frac{\partial \sigma_{ri}}{\partial \alpha} \right)^T \left(\frac{\partial \sigma_{sj}}{\partial \alpha} \right) \right] k^T
 \end{aligned}$$

where, in this derivation, we have made use of the identity

$$\lambda_{ri} E_{ri} W(\alpha) \lambda_{sj} E_{sj} = \lambda_{ri} w_{rs}(\alpha) \lambda_{sj} E_{ij}.$$

Finally,

$$\begin{aligned}
& kXW(\alpha) \delta \sum(\alpha; k) W(\alpha) X^T \alpha^T \\
&= \text{Tr}[\delta \sum(\alpha; k)] [W(\alpha) X^T k^T \alpha X W(\alpha)] \\
&= \text{Tr} \sum_{i=1}^n \sum_{j=1}^n [\delta \sigma_{ij}(\alpha; k) E_{ij}] [W(\alpha) X^T k^T \alpha X W(\alpha)] \\
&= \sum_{i=1}^n \sum_{j=1}^n \delta \sigma_{ij}(\alpha; k) \text{Tr}(kXW(\alpha) E_{ij}^T W(\alpha) X^T \alpha^T) \\
&= \sum_{i=1}^n \sum_{j=1}^n k \left(\frac{\partial \sigma_{ij}}{\partial \alpha} \right)^T (\alpha XW(\alpha) E_{ij} W(\alpha) X^T k^T) \\
&= k \left[\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\partial \sigma_{ij}}{\partial \alpha} \right)^T (\alpha XW(\alpha) E_{ij} W(\alpha) X^T) \right] k^T \\
&= k \left[\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \sum_{s=1}^n w_{sj}(\alpha) \left(\frac{\partial \sigma_{ij}}{\partial \alpha} \right)^T X(s) \right] k^T
\end{aligned}$$

where we have used the auxiliary identity

$$\alpha XW(\alpha) E_{ij} W(\alpha) X^T = \sum_{s=1}^n \mu_i(\alpha) w_{sj}(\alpha) X(s) \quad .$$

Bringing these results together we find that

$$\delta^2 Q(\alpha; k) = 2kB_{\alpha} k^T$$

where the matrix B_{α} is given by

$$\begin{aligned}
B_{\alpha} = & \sum_{i=1}^n \sum_{j=1}^n X^T(i) w_{ij}(\alpha) X(j) - \sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \Sigma_{ij} \\
& + \sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \mu_j(\alpha) \sum_{r=1}^n \sum_{s=1}^n w_{rs}(\alpha) \left(\frac{\partial \sigma_{ri}}{\partial \alpha} \right)^T \left(\frac{\partial \sigma_{sj}}{\partial \alpha} \right) \\
& - 2 \sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \sum_{s=1}^n w_{sj}(\alpha) \left(\frac{\partial \sigma_{ij}}{\partial \alpha} \right)^T X(s) .
\end{aligned}$$

Combining this expression for $\delta^2 Q(\alpha; k)$ with Theorem 3.5.5 we have in effect established the following result, namely,

Theorem 4.5.6. If $\delta Q(\alpha_0; k) = 0$ identically in k , i.e., if α_0 is a critical point of $Q(\alpha) = \alpha X W(\alpha) X^T \alpha^T = d(\alpha) W(\alpha) d^T(\alpha)$ where

$$W(\alpha) = [w_{ij}(\alpha)] = [\sigma_{ij}(\alpha)]^{-1} = \Sigma_d^{-1}(\alpha)$$

and

$$\sigma_{ij}(\alpha) = \alpha \Sigma_{ij} \alpha^T$$

then α_0 is a minimum of $Q(\alpha)$ provided the matrix B_{α_0} is positive definite where $B_{\alpha_0} =$

$$\begin{aligned}
& \sum_{i=1}^n \sum_{j=1}^n \left[X^T(i) w_{ij}(\alpha_0) X(j) + \mu_i(\alpha_0) \mu_j(\alpha_0) \sum_{r=1}^n \sum_{s=1}^n w_{rs}(\alpha_0) \left(\frac{\partial \sigma_{ri}}{\partial \alpha_0} \right)^T \left(\frac{\partial \sigma_{sj}}{\partial \alpha_0} \right) \right. \\
& \quad \left. - \mu_i(\alpha_0) \mu_j(\alpha_0) \Sigma_{ij} - 2 \mu_i(\alpha_0) \sum_{s=1}^n w_{sj}(\alpha_0) \left(\frac{\partial \sigma_{ij}}{\partial \alpha_0} \right)^T X(s) \right]
\end{aligned}$$

and $\partial \sigma_{ij} / \partial \alpha_0$ is the Dwyer-Macphail derivative of $\sigma_{ij}(\alpha)$ evaluated at $\alpha = \alpha_0$.

The remarks made previously for the EVE model when $W_*(\theta)$ was assumed to be a diagonal matrix remain relevant in the general case of non-diagonal $W_*(\theta)$; and consequently a similar result holds for the EVE model which we state here for the sake of completeness, the same proof as given above for the EV model when $W(\alpha)$ is non-diagonal being valid:

Theorem 4.5.7. If $\delta Q_*(\theta_0; h) = 0$ identically in h , i.e., if θ_0 is a critical point of $Q_*(\theta) = \theta Z W_*(\theta) Z^T \theta^T = d_*(\theta) W_*(\theta) d_*^T(\theta)$ where

$$W_*(\theta) = [w_{ij}^*(\theta)] = [\sigma_{ij}^*(\theta)]^{-1} = \Sigma_{d_*}^{-1}(\theta)$$

and

$$\sigma_{ij}^*(\theta) = \sigma_{ij}^*(\xi) + \Gamma_{ij}^* \theta^T + \theta \Sigma_{ij}^* \theta^T,$$

then θ_0 is a minimum of $Q_*(\theta)$ provided the matrix $B_{\theta_0}^*$ is positive definite where

$$B_{\theta_0}^* = \sum_{i=1}^n \sum_{j=1}^n \left[Z^T(i) w_{ij}^*(\theta_0) Z(j) + \mu_i^*(\theta_0) \mu_j^*(\theta_0) \sum_{r=1}^n \sum_{s=1}^n w_{rs}^*(\theta_0) \left(\frac{\partial \sigma_{ri}^*}{\partial \theta_0} \right)^T \left(\frac{\partial \sigma_{sj}^*}{\partial \theta_0} \right) \right. \\ \left. - \mu_i^*(\theta_0) \mu_j^*(\theta_0) \Sigma_{ij}^* - 2\mu_i^*(\theta_0) \sum_{s=1}^n w_{sj}^*(\theta_0) \left(\frac{\partial \sigma_{ij}^*}{\partial \theta_0} \right)^T Z(s) \right]$$

and $\partial \sigma_{ij}^* / \partial \theta_0$ is the Dwyer-Macphail matrix derivative of $\sigma_{ij}^*(\theta)$ evaluated at $\theta = \theta_0$.

Finally, we mention briefly a third type of regularity consideration associated with the "identification" problem in EV and EVE statistical

models. The identification problem can be illustrated by the following simple example. Suppose

$$S(i) = a + bR(i)$$

where, for $1 \leq i \leq n$, we observe

$$X(i) = R(i) + \delta(i) \quad \text{and} \quad Y(i) = S(i) + \eta(i)$$

and $R(i)$ and the errors $\delta(i)$ and $\eta(i)$ are jointly normally distributed and uncorrelated with

$$E(R(i), \delta(i), \eta(i)) = (\mu, 0, 0)$$

and

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma_\delta^2 & 0 \\ 0 & 0 & \sigma_\eta^2 \end{bmatrix} .$$

In this case we find that the $Z(i) = (X(i), Y(i))$ are normally distributed with mean

$$EZ(i) = (\mu, a+b\mu) = E(X(i), Y(i))$$

and covariance

$$\Sigma_{ii} = \begin{bmatrix} \sigma^2 + \sigma_\delta^2 & b\sigma^2 \\ b\sigma^2 & b^2\sigma^2 + \sigma_\eta^2 \end{bmatrix} = V .$$

Thus if

$$Z = (Z(1), Z(2), \dots, Z(n))$$

the joint distribution function of Z is given by

$$\begin{aligned} F(z_1, z_2, \dots, z_n | a, b, \mu, \sigma^2, \sigma_\delta^2, \sigma_\eta^2) \\ = \Pr(Z(1) < z_1, Z(2) < z_2, \dots, Z(n) < z_n) \\ = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |V|^{-\frac{n}{2}} \int_{-\infty}^z \exp \sum_{i=1}^n (X(i) - \mu, Y(i) - a - b\mu) V^{-1} (X(i) - \mu, Y(i) - a - b\mu)^T dz; \end{aligned}$$

where this integral is understood to be a multiple integral over the indicated interval in $2n$ -dimensional Euclidean space.

Now let us consider the two specific models

$$M_1: a_1 = 9; b_1 = 1; \mu_1 = 1; \sigma_1^2 = 1; \sigma_1^2(\delta) = \frac{1}{9}; \sigma_1^2(\eta) = 9$$

and

$$M_2: a_2 = 1; b_2 = 9; \mu_2 = 1; \sigma_2^2 = \frac{1}{9}; \sigma_2^2(\delta) = 1; \sigma_2^2(\eta) = 1;$$

we find that

$$F_1(z_1, z_2, \dots, z_n) \equiv F_2(z_1, z_2, \dots, z_n)$$

where

$$F_j(z_1, z_2, \dots, z_n) = F(z_1, z_2, \dots, z_n | a_j, b_j, \mu_j, \sigma_j^2, \sigma_j^2(\delta), \sigma_j^2(\eta))$$

Thus we see that two different models can generate observations

$Z(i) = (X(i), Y(i))$ which have identical distribution functions.

Now consider an estimator $\hat{b}$ of b ; as an estimator $\hat{b}$ is a function of the observations, say

$$\hat{b} = \hat{b}(Z(1), Z(2), \dots, Z(n)) ;$$

and so its probability distribution is determined by that of the $Z(k)$'s and consequently the distribution function of b is compatible with two distinct values of b , namely $b_1 = 1$ and $b_2 = 9$. Since a probability limit is unique, it follows that $\hat{b}$ cannot converge to b in probability and therefore without additional prior information no estimator of b based only on the observations alone can be even a weakly consistent estimator. Since weak consistency may be taken as a minimal type property that an estimator should possess, we find that in order to obtain weakly consistent estimators in this situation requires additional prior information about the underlying stochastic structure.

This, of course, is a very rough statement of the "identification" problem when the variances and covariances of the disturbances of an EV model are unknown and normally distributed and we refer the interested reader to [78,83,121] where important results and other references on this problem can be found. This example, however, does point up the necessity of using prior information about basic parameters of a statistical model, at least when the model disturbances are normally distributed, in order to obtain useful estimators of the unknown structural parameters of the model. An analysis of the identification problem for linear structural relationships having errors-in-variables and errors-in-equations apparently has not been made, although prior knowledge of the variances and covariances of the variable and equation

errors and the absence of linear relationships among these disturbances would appear to be a sufficient condition for identifiability of the parameters of the EVE model.

This concludes our discussion of regularity conditions associated with GLS-estimators of the unknown parameters of the EV and EVE statistical models. In the next chapter, assuming that the variable and equation errors are normally distributed, we find that the maximum likelihood estimators of these parameters are identical to the GLS-estimators studied in this chapter.

CHAPTER 5

ESTIMATION IN MODIFIED EV AND EVE STATISTICAL MODELS

5.1. INTRODUCTION

In the final chapter of this dissertation we examine the estimation problem for modified EV and EVE statistical models in which additional assumptions are made about the probability distributions of the random elements of these two stochastic structures. One of the strongest additional specifications that we make is to assume that the equation or variable disturbances are jointly normally distributed. When the model errors are normally distributed, the maximum likelihood estimation procedure is readily applicable, of course, and we derive the generally non-linear ML-normal equation system which implicitly define the ML-estimators of the unknown parameters of these modified EV and EVE statistical models.

Writing down the likelihood function for the "observables" $Z(i) = (X(i), Y(i))$ of the EVE model, we find that this functional depends not only on the parameters $\theta = (\alpha, \beta)$ but also on the unknown "incidental" parameters $E(Z(i)) = (EX(i), EY(i)) = (r(i), s(i))$ as well, where $s(i) = ES(i)$ for $1 \leq i \leq n$; and the same situation is found for the EV model too where the likelihood function is based on the "observables" $X(i)$ for $1 \leq i \leq n$. Hence we must find extreme points of this likelihood which are subject to the restraints imposed upon the

incidental parameters of the model by the structural relationship existing between the variables of the model.

It turns out, however, that this rather complex constrained maximization problem which represents the application of the maximum likelihood estimation procedure to the likelihood function of the observable variables of the EV or EVE models leads us to the generalized least squares estimation procedure examined in Chapter 4 when the joint distribution of the variable and equation errors was left unspecified.

Besides investigating the maximum likelihood estimation procedure when the variable or equation disturbances of the EV and EVE statistical models are assumed to be jointly normally distributed, we also examine the generalized least squares estimation procedure in certain EV models in which the variable errors, although not assumed to be normally distributed, are assumed to have rather simple covariance matrices. The particular covariance matrices selected for consideration here are those that had been assumed by other authors in published works on the estimation problem in the EV statistical model. The celebrated Gauss-Markoff-Aitken estimators for the explicit EV model, for example, in which only the dependent variable is assumed to be in error can be obtained as a special case of our more general results.

Efforts to extend their work to other EV models in which several variables can be in error have occupied the attention of many research workers in mathematical statistics and various results in this direction

have been published, some of very recent origin. In these works, however, specific covariance matrices of a very simple nature usually are assumed among the model errors, and the derivations given for the normal equations defining the estimators depend for their validity on the simplified nature of the special covariance matrices assumed by the authors. Thus, for the first time, their solutions to the estimation problem in the EV model are seen to be particular cases of our general solution.

Another modified EV statistical model associated with linear regression analysis when all variables are subject to errors had been considered previously by two other authors. K. Norton in [113], working with an explicit structural relationship, did not consider the general estimation problem for the EV model related to regression theory when non-observable variables are present, but concentrated on obtaining the normal equations for modified GLS-estimators based on a second order Taylor's series expansion of the least squares functional about a fixed, predetermined estimate of the unknown parameters of the structure relation. Miss Parel in [114] gave a general matrix presentation of Norton's modified GLS-estimation procedure, and she also obtained the GLS-normal equations for the linear regression model whose variables are subject to errors when such errors are assumed to be intercorrelated and serially correlated. Again, her results can be obtained as a special case of our results for the explicit EV model.

We now go on to derive for the first time maximum likelihood estimators of the unknown parameters of the EV and EVE statistical models

when the variable and equation disturbances are assumed to be jointly normally distributed and to be both intercorrelated and serially correlated as well; and we show that the maximum likelihood estimators of the structural parameters of the EV and EVE statistical models are identical to the GLS-estimators of these parameters studied in Chapter 4.

5.2. MAXIMUM LIKELIHOOD ESTIMATORS FOR THE EV MODEL

Consider the EV model $\alpha \cdot r^T(i) = 0$ where we observe $X(i) = r(i) + \delta(i)$ for $1 \leq i \leq n$ and α , $r(i)$ and $\delta(i)$ are p -dimensional vectors. Let

$$r = (r(1), r(2), \dots, r(n)),$$

$$\delta = (\delta(1), \delta(2), \dots, \delta(n)),$$

and

$$X = (X(1), X(2), \dots, X(n)) = r + \delta ;$$

we assume that the pn -dimensional random vector δ is normally distributed about 0 with the known, non-singular covariance matrix Σ_δ . Setting $W = \Sigma_\delta^{-1}$ and

$$\begin{aligned} \|X-r\|_W &= (X-r)W(X-r)^T \\ &= \sum_{i=1}^n \sum_{j=1}^n X(i)W_{ij}X^T(j) \\ &= \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^p \sum_{s=1}^p X_r(i)w_{ij}^{rs}X_s(j) , \end{aligned}$$

an explicit expression for the joint density function of the random vec-

for X is given by,

$$f(X; \alpha, r) = \left(\frac{1}{2\pi} \right)^{np/2} |\Sigma_0|^{-\frac{1}{2}} e^{-\frac{1}{2} \|X-r\|^2} \Sigma_0^{-1}$$

or

$$f(X; \alpha, r) = \left(\frac{1}{2\pi} \right)^{np/2} |W|^{-\frac{1}{2}} e^{-\frac{1}{2} (X-r)W(X-r)^T},$$

where the density function f depends on the unknown parameter α implicitly through the constraints that the n -vector components of r satisfy, namely $\alpha \cdot r^T(i) = 0$ for $1 \leq i \leq n$.

Using a stepwise maximization procedure, we first maximize $f(X; \alpha, r)$ with respect to r for fixed but arbitrary α ; thus we seek $\hat{r} = \hat{r}_\alpha(x)$ such that $f(X; \alpha, \hat{r}) \geq f(X; \alpha, r)$ for every r subject to the constraints $\alpha \cdot r^T(i) = 0$ for $1 \leq i \leq n$. Introducing Lagrangian multipliers $\psi_1, \psi_2, \dots, \psi_n$ and using the likelihood

$$\ell(r|\alpha; X) = \ln f(X; r, \alpha),$$

we form the Lagrangian function for this problem, namely

$$L_\psi(r|\alpha, X) = \ell(r|\alpha, X) + \psi \cdot \Lambda^T(r|\alpha)$$

where

$$\psi = (\psi_1, \psi_2, \dots, \psi_n),$$

$$\Lambda(r|\alpha) = (\alpha \cdot r^T(1), \alpha \cdot r^T(2), \dots, \alpha \cdot r^T(n)),$$

and

$$\ln f(X; \alpha, r) = - \left(\frac{np}{2} \right) \ln 2\pi + \frac{1}{2} \ln |W| - \frac{1}{2} (X-r)W(X-r)^T.$$

By a direct derivation, we see that the Dwyer-Macphail matrix derivative $\partial L / \partial r$ of the Lagrangian $L_\psi(r|\alpha, X)$ is given by

$$\frac{\partial L}{\partial r} = -2(X-r)W + (\psi_1\alpha, \psi_2\alpha, \dots, \psi_n\alpha).$$

Hence, according to Theorem 3.5.3, a necessary condition that $\hat{r}$ be a critical point of $L_\psi(r|\alpha, X)$ is that $\partial L / \partial r]_{r=\hat{r}} = 0$ or that

$$2(X-\hat{r})W = (\psi_1\alpha, \psi_2\alpha, \dots, \psi_n\alpha).$$

But this necessary condition $\partial L / \partial r]_{r=\hat{r}} = 0$ is equivalent to the Frechet-Gateaux differential $\delta L_\psi(\hat{r}; k|\alpha, X) = (\partial L_\psi / \partial \hat{r}) k^T$ vanishing for every non-zero vector k ; in particular, taking $k = \sum_{\mathcal{G}} k_{\star}^T$ where $k_{\star}$ is a fixed but arbitrary non-zero vector, we find that

$$\begin{aligned} 2(X-\hat{r})k_{\star}^T &= 2(X-\hat{r})(W W^{-1})k_{\star}^T \\ &= 2(X-\hat{r})W(\sum_{\mathcal{G}} k_{\star}^T) \\ &= \sum_{r=1}^n \sum_{j=1}^n \psi_r \alpha \sum_{r,j} k_{\star}^T(j) \end{aligned}$$

where $\sum_{r,j} = E(\delta^T(r)\delta(j))$ is a submatrix of $\sum_{\mathcal{G}}$ and $k_{\star}(j)$ is the j -th vector component of $k_{\star}$. We rewrite this identity in the form

$$2 \sum_{j=1}^n \hat{r}(j) k_{\star}^T(j) = 2 \sum_{j=1}^n X(j) k_{\star}^T(j) - \sum_{j=1}^n \left(\sum_{r=1}^n \psi_r \alpha \sum_{r,j} \right) k_{\star}^T(j);$$

and in this last expression we set $k_*(j) = \alpha$ and $k_*(i) = 0$ for $i \neq j$ obtaining

$$2\hat{f}(j)\alpha^T = 2X(j)\alpha^T - \sum_{r=1}^n \psi_r \alpha \sum_{rj} \alpha^T = 0$$

since $\hat{f}(j)$ was determined so as to satisfy the restraint $\hat{f}^T(j) = 0$ for $1 \leq j \leq n$. Thus

$$\sum_{r=1}^n (\alpha \sum_{rj} \alpha^T) \psi_r = 2\alpha X^T(j)$$

for $1 \leq j \leq n$; and we have obtained a system of linear equations in the n -unknowns ψ_r . Provided

$$W(\alpha) = [w_{ij}(\alpha)] = [\sigma_{ij}(\alpha)]^{-1} = [\alpha \sum_{ij} \alpha^T]^{-1}$$

exists we can solve this system of equations for the ψ_r uniquely as

$$\psi^T = 2W(\alpha) X^T \alpha^T$$

where

$$X = [X^T(1), X^T(2), \dots, X^T(n)]$$

Thus

$$\begin{aligned} \psi_i &= 2 \sum_{j=1}^n w_{ji}(\alpha) X(j) \alpha^T \\ &= 2 \sum_{j=1}^n w_{ij}(\alpha) X(j) \alpha^T \end{aligned}$$

since $W(\alpha) = W^T(\alpha)$ or $w_{ij}(\alpha) = w_{ji}(\alpha)$; and substituting this expression

for ψ_r into the expression given above for $\hat{f}^T_*$ we find that

$$\sum_{j=1}^n \hat{r}(j) k_*^T(j) = \sum_{j=1}^n \left(X(j) - \sum_{r=1}^n \alpha \left(\sum_{s=1}^n w_{rs}(\alpha) \alpha X^T(s) \right) \Sigma_{rj} \right) k_*^T(j)$$

from which we conclude that for $1 \leq j \leq n$

$$\hat{r}(j) = X(j) - \sum_{r=1}^n \alpha \left(\sum_{s=1}^n w_{rs}(\alpha) \alpha X^T(s) \right) \Sigma_{rj} .$$

Hence as a result of the first stage of our stepwise maximization procedure, for fixed but arbitrary non-zero α , we have obtained an explicit expression for the ML-estimators of the "incidental" parameter "r" of the EV statistical model subject to the constraints $\alpha \cdot r^T(i) = 0$ for $1 \leq i \leq n$. We now seek the unconstrained ML-estimator of α which we obtain by maximizing

$$f^*(X; \alpha) = \left(\frac{1}{2\pi} \right)^{np/2} |W|^{1/2} e^{-\frac{1}{2} (X - \hat{r}(\alpha)) W (X - \hat{r}(\alpha))^T} = f(X; \alpha, \hat{r}(\alpha)) .$$

Thus we must express $(X - \hat{r}(\alpha)) W (X - \hat{r}(\alpha))^T$ as an explicit function of α .

This can be accomplished readily by considering the necessary condition

$$\delta L_{\psi}(\hat{r}; k | \alpha, X) = \left(\frac{\partial L_{\psi}}{\partial \hat{r}} \right) k^T = 0$$

or

$$2(X - \hat{r}) W k^T = \sum_{i=1}^n \psi_i \alpha k^T(i)$$

which holds identically for $k \neq 0$, and in this identity taking $k = X - \hat{r}$.

Then we see that

$$\begin{aligned}
2(X - \hat{A})^T W (X - \hat{A}) &= \sum_{i=1}^n \psi_i \alpha (X^T(i) - \hat{A}^T(i)) \\
&= \sum_{i=1}^n \psi_i (\alpha \cdot X^T(i) - \alpha \cdot \hat{A}^T(i)) \\
&= \sum_{i=1}^n \alpha \cdot X^T(i) \psi_i \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) w_{ij}(\alpha) X(j) \alpha^T .
\end{aligned}$$

Consequently we find that

$$f^*(X; \alpha) = \left(\frac{1}{2\pi} \right)^{np/2} |W|^{1/2} e^{-\frac{1}{4} Q(\alpha)}$$

where

$$Q(\alpha) = \sum_{i=1}^n \sum_{j=1}^n \alpha X^T(i) w_{ij}(\alpha) X(j) \alpha^T .$$

Now $Q(\alpha)$ is just the functional for the GLS-estimation procedure which we studied in Chapter 4; and since the ML-estimator $\hat{\alpha}$ or α is obtained by maximizing $f^*(X; \alpha)$ or, equivalently, by minimizing the likelihood

$$l_*(\alpha|X) = \ln f^*(X; \alpha)$$

or

$$l_*(\alpha|X) = - \left(\frac{np}{2} \right) \ln 2\pi + \frac{1}{2} \ln |W| - \frac{1}{4} Q(\alpha) ,$$

we see that $\hat{\alpha}$ minimizes $Q(\alpha)$ and thus is identical with the GLS-estimator

of α previously studied.

Combining these observations into a single statement, we see that we have established the following result, namely,

Theorem 5.2.1. If the errors $\delta(i)$ of the EV statistical model are jointly normally distributed about 0 with known, non-singular covariance matrix $\Sigma_\delta = [\Sigma_{ij}]$ where $\Sigma_{ij} = E\delta^T(i)\delta(j)$ for $1 \leq i, j \leq n$, then maximum likelihood estimators $\hat{A}(k)$ of the incidental parameters $r(k)$ for $1 \leq k \leq n$ and $\hat{\alpha}$ of the structural parameter α of the EV model are given by

$$\hat{A}(k) = X(k) - \sum_{r=1}^n \hat{\alpha} \left(\sum_{s=1}^n w_{rs}(\hat{\alpha}) \hat{A} X^T(s) \right) \Sigma_{rk}$$

and any non-zero $\hat{\alpha}$ which is a minimum of the functional

$$Q(\alpha) = \sum_{r=1}^n \sum_{j=1}^n \alpha X^T(i) w_{ij}(\alpha) X^T(j) \alpha^T$$

where

$$W(\alpha) = [w_{ij}(\alpha)] = [\alpha \Sigma_{ij} \alpha^T]^{-1}$$

Hence the maximum likelihood estimator $\hat{\alpha}$ is identical to the GLS-estimator of the structural parameters of the EV model.

5.3. MAXIMUM LIKELIHOOD ESTIMATORS FOR THE EVE MODEL

Once the likelihood of the observations $Z(i) = (X(i), Y(i))$ for $1 \leq i \leq n$ is obtained, we find that the ML-analysis of the EVE model is **formally the** same as the ML-analysis of the EV model given in Section 5.2. To establish the joint distribution function of the observations, let

us recall that in the EVE model

$$\theta \cdot U^T(i) = \alpha \cdot r^T(i) + \beta \cdot S^T(i) = \xi(i)$$

for $1 \leq i \leq n$ where $U(i) = (r(i), S(i))$ and $\theta = (\alpha, \beta)$; and instead of observing $r(i)$ and $S(i)$ directly we observe

$$X(i) = r(i) + \delta(i) \quad \text{and} \quad Y(i) = S(i) + \eta(i) .$$

Set $V(i) = (\delta(i), \eta(i))$ and $Z(i) = (X(i), Y(i))$ and consider the random vectors

$$U = (U(1), U(2), \dots, U(n)) ,$$

$$V = (V(1), V(2), \dots, V(n)) ,$$

and

$$Z = (Z(1), Z(2), \dots, Z(n)) = U + V .$$

Now when $S(i)$, $\delta(i)$ and $\eta(i)$ are jointly normally distributed, V is normally distributed about 0 with covariance matrix $\Sigma_V = [\Sigma_{ij}^{(V)}]$ where

$$\Sigma_{ij}^{(V)} = \begin{bmatrix} \Sigma_{ij}^{(\delta)} & \Sigma_{ij}^{(\delta, \eta)} \\ \Sigma_{ij}^{(\eta, \delta)} & \Sigma_{ij}^{(\eta)} \end{bmatrix} = E(V^T(i) V(j))$$

and U is normally distributed about $u = (u(1), u(2), \dots, u(n))$ where $u(i) = (r(i), s(i))$ with $s(i) = ES(i)$ and with covariance matrix

$$\Sigma_U = [\Sigma_{ij}^{(U)}] \text{ where}$$

$$\Sigma_{ij}^{(U)} = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{ij}^{(S)} \end{bmatrix} = E(U(i) - u(i)) (U(i) - u(i))^T \quad .$$

Consequently, since Z is a linear combination of two jointly normally distributed variables, Z must be normally distributed; moreover,

$$z = E(Z) = (z(1), z(2), \dots, z(n)) = E(U)$$

and

$$\Sigma_Z = \Sigma_U + \Sigma_{(U,V)} + \Sigma_{(V,U)} + \Sigma_V = E(Z - z)(Z - z)^T$$

or

$$[\Sigma_{ij}^{(Z)}] = [\Sigma_{ij}^{(U)} + \Sigma_{ij}^{(U,V)} + \Sigma_{ij}^{(V,U)} + \Sigma_{ij}^{(V)}]$$

where

$$\Sigma_{ij}^{(Z)} = E((Z(i) - z(i))(Z(j) - z(j)))$$

$$= \begin{bmatrix} \Sigma_{ij}^{(\delta)} & \Sigma_{ij}^{(\delta, \eta)} \\ \Sigma_{ij}^{(S, \delta)} + \Sigma_{ij}^{(\delta, S)} + \Sigma_{ij}^{(\eta, \delta)} & \Sigma_{ij}^{(S)} + \Sigma_{ij}^{(S, \eta)} + \Sigma_{ij}^{(\eta, S)} + \Sigma_{ij}^{(\eta)} \end{bmatrix} \quad .$$

In other words, the joint density function of Z is given by

$$f(Z; \theta, z) = \left(\frac{1}{2\pi} \right)^{npq/2} |W_*|^{1/2} e^{-\frac{1}{2}(Z-z)^T W_* (Z-z)} ;$$

here $W_* = \sum_Z^{-1}$ and $z = (z(1), z(2), \dots, z(n)) = E(Z)$ where the

$z(i) = (r(i), s(i))$ with $s(i) = ES(i)$ are subject to the constraints

$$\theta \cdot z^T(i) = \alpha \cdot r^T(i) + \beta \cdot s^T(i) = 0$$

for $1 \leq i \leq n$.

According to the maximum likelihood estimation procedure, we seek

$\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ and $\hat{z}$ such that

$$f(Z; \hat{\theta}, \hat{z}) \geq f(Z; \theta, z)$$

subject to the restraints $\theta \cdot z^T(i) = 0$ for $1 \leq i \leq n$. Thus the ML-estimation procedure applied to the EVE statistical model turns out to be formally the same as the ML-estimation procedure applied to the EV model; and upon identifying θ with α , z with r , W_* with W and $\sum_Z$ with $\sum_{\delta}$ while recalling the identities $\sigma_{ij}^{(\xi)} = \beta \sum_{ij}^{(S)} \beta^T$,

$$\beta \sum_{ij}^{(S, \delta)} \alpha^T = \Gamma_{ij}^{(\xi, \delta)} \alpha^T \quad \text{and} \quad \beta \sum_{ij}^{(S, \eta)} \beta^T = \Gamma_{ij}^{(\xi, \eta)} \beta^T$$

established in Chapter 4, by repeating the derivation given in Section 5.2, the following result is obtained:

Theorem 5.3.1. If the variables and disturbances of the EVE statistical model are jointly normally distributed with known, covariance matrix, while the variable and equation errors have 0 means, then the

maximum likelihood estimators $\hat{z}(k) = (\hat{r}(k), \hat{s}(k))$ of the incidental parameters $z(k) = (r(k), s(k))$ for $1 \leq k \leq n$ and $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ of the structural parameter $\theta = (\alpha, \beta)$ of the EVE model are given by

$$(\hat{r}(k), \hat{s}(k)) = (X(k), Y(k))$$

$$- \sum_{r=1}^n (\hat{\alpha}, \hat{\beta}) \left[\sum_{s=1}^n w_{rs}^*(\hat{\alpha}, \hat{\beta}) (\hat{\alpha} \cdot X^T(s) + \hat{\beta} \cdot Y^T(s)) \right] \Sigma_{rk}^{(Z)}$$

and any non-zero $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ which is a minimum of the functional

$$Q_*(\theta) = \sum_{i=1}^n \sum_{j=1}^n \theta Z^T(i) w_{ij}^*(\theta) Z^T(j) \theta^T$$

where $Z(i) = (X(i), Y(i))$ for $1 \leq i \leq n$,

$$W_*(\theta) = [w_{ij}^*(\theta)] = [\sigma_{ij}^*(\theta)]^{-1},$$

and

$$\sigma_{ij}^*(\theta) = \sigma_{ij}^{(\xi)} + \Gamma_{ij}^* \theta^T + \theta \Sigma_{ij}^* \theta^T.$$

Hence the maximum likelihood estimator $\hat{\theta}$ is identical to the GLS-estimator of the structural parameter θ of the EVE model.

5.4. CONNECTIONS WITH OTHER RESEARCH ON THE ESTIMATION PROBLEM IN THE EV AND EVE MODELS

We have seen in Sections 5.2 and 5.3, when the disturbances of the EV and EVE models are assumed to be jointly normally distributed, that the maximum likelihood estimation procedure based on the joint distribution function of the observations $X(i)$ or $Z(i) = (X(i), Y(i))$, leads

us to ML-estimators of the structural parameters α or $\theta = (\alpha, \beta)$ which are identical with the GLS-estimators for these parameters of the EV and EVE models developed in Chapter 4. In this section we go on to explore the connections between this work and the results of other authors who studied the estimation problem for the EV statistical model by means of the least squares principle or the maximum likelihood estimation procedure, but generally under more simplified assumptions than we have made regarding the intercorrelations and serial correlations of the model's disturbances or errors.

For the error-in-variable model we relate the results of the following authors to our developments:

- The Gauss-Markoff-Aitken EV model specification;
- The Kummel-Lindley EV model specification;
- The Koopmans-Tintner EV model specification;
- The Villegas EV model specification;
- The Parel EV model specification; and
- The Deming-Norton-Parel EV model specification,

the latter two EV model specifications being associated with the linear regression model when non-observable variables are present.

Turning our attention to the EVE statistical model, since it can be interpreted as an EV model in which additional stochastic elements are present—namely the random scalar variables $\{S_j(k)\}_{j=1}^q$ as well as the sure scalar variables $\{r_i(k)\}_{i=1}^p$, the additional errors $\{\eta_j(k)\}_{j=1}^q$ associated with these random variables, and the equation disturbances $\{\xi(k)\}_{k=1}^n$ of the structural relationship—the EV model can be considered as a modified EVE model. But except in so far as the EV model specifi-

cations that we have mentioned can be interpreted as modified EVE models and thus the results obtained by other authors for these EV model specifications introduced as special cases of modified EVE models, for the general EVE model in which these additional stochastic elements are present, this author is not aware of any results having been obtained earlier by other authors for the general EVE statistical model.

The Parel EV Model Specification

First we discuss the Parel EV model specification and the Deming-Norton-Parel modification of the GLS-estimation procedure for this specification. In [114] Miss Parel (essentially) considers an explicit structural relation

$$x(i) = \alpha_0 + \alpha_1 a_1(i) + \dots + \alpha_s a_s(i) + \xi(i)$$

in which, for $1 \leq i \leq n$, the $a_1(i), a_2(i), \dots, a_s(i)$ are considered to be fixed, non-observable sure scalar variates and the $\xi(i)$ and $x(i)$ are random variables having $E\xi(i) = 0$ and $Ex(i) = a(i)$. Rather than observing the non-random scalar variates $a_k(i)$ or the random variables $x(i)$ directly, she assumes that one observes

$$x^0(i) = x(i) + \gamma(i) \quad \text{and} \quad a_k^0(i) = a_k(i) + \delta_k(i) \quad \text{for } 1 \leq k \leq s.$$

In addition Miss Parel assumes that the $\delta_k(i)$, $\gamma(i)$ and $\xi(i)$ can be intercorrelated and serially correlated. To this author's knowledge her work stands as one of the most general approaches to the estimation prob-

lem for the explicit EV statistical model .

Her results can be obtained from those on GLS-estimation for the explicit EV model that we have presented as follows: Consider

$$E\xi(i) = \alpha_0 + \alpha_1 a_1(i) + \dots + \alpha_s a_s(i) - a(i) = 0 ,$$

an implicit structural relationship in which only sure variables appear, or equivalently, the explicit structural relationship

$$a(i) = \alpha_0 + \alpha_1 a_1(i) + \dots + \alpha_s a_s(i) .$$

Then we observe

$$x^0(i) = x(i) + \gamma(i) = a(i) + \gamma(i) + \xi(i) = a(i) + \delta(i)$$

where we have set $\delta(i) = \gamma(i) + \xi(i)$ for $1 \leq i \leq n$. In addition we observe $a_k^0(i) = a_k(i) + \delta_k(i)$ for $1 \leq k \leq s$. But this however is just an explicit EV model in which the errors of the dependent variable have the form $\delta(i) = \gamma(i) + \xi(i)$. Thus the normal equations defining the GLS-estimators of the unknown parameters of the linear regression model when errors-in-variables are present can be written down at once from the general results that we have presented in Section 4.2 for the explicit EV model.

The Deming-Norton-Parel EV Model Specification

Besides obtaining these GLS-estimation equations for the unknown parameters of the linear regression model when all variables are subject

to error, Miss Parel also studied two other important questions connected with this estimation problem. She showed that when "constellations" of observations are available, GLS-estimations can be obtained even though the variances and covariances of the error disturbances are unknown. She accomplished this by using "constellation means" to form the GLS-likelihood instead of the individual multivariate observations and substituting sample estimates obtained (from these constellations) for the covariance matrices of the transformed disturbances corresponding to the constellation means.

Next Miss Parel gave an iterative computational procedure for obtaining an approximate solution to these GLS-normal equations, she studied the convergence properties of this approximation technique, and then applied these results to several examples. Basically, this iterative computational technique is an adaptation of a procedure advanced by Deming [27] and applied earlier by Norton [113] to this error-in-variable model. It involves expanding the GLS-likelihood function in a second degree series expansion, which Miss Parel provided in a form which featured the second order matrix derivatives that she had introduced for this purpose. Then taking this second degree Taylor's sum as a modified GLS-likelihood function, its minimum value is taken as an approximation to the GLS-estimator. By forming this second degree expansion in subsequent stages about the last modified GLS-estimator so obtained, Miss Parel showed that under certain regularity conditions this sequence of estimators converges to the GLS-estimator.

We have discussed the GLS-normal equations for Miss Parel's specification of the EV model and indicated how they can be obtained as a special case of our GLS-normal equations for the explicit EV model presented in Chapter 4; and we have briefly summarized how the Deming-Norton-Parel modified GLS-normal equations lead to an approximate solution which converges under certain regularity conditions to the GLS-estimators for this EV model. We now go on to show how results established by other authors for the error-in-variable or EV model can be obtained as special cases of our analysis too.

Gauss-Markoff-Aitken EV Model Specification

The reader may find it interesting to know that our normal equations $A_{\alpha_*}^T \alpha_*^T = 0$, defining the GLS-estimators of the parameters of the explicit EV statistical model in which all variables are subject to errors that may be correlated, generalize the celebrated Gauss-Markoff least squares estimators as extended by Aitken [2] and others[118]. Thus consider the stochastic equation

$$y(i) = \alpha_1 + \alpha_2 x_2(i) + \dots + \alpha_{p-1} x_{p-1}(i) + e(i)$$

where the $x_j(i)$'s for $2 \leq j \leq p-1$ are fixed, known scalar variates and the $e(i)$'s for $1 \leq i \leq n$ are random variables with $Ee(i) = 0$ and known variance-covariance matrix $\Sigma_e = Ee^T e$. We may write this model in the form

$$r(i) = Ey(i) = \alpha_1 + \alpha_2 x_2(i) + \dots + \alpha_{p-1} x_{p-1}(i)$$

where we observe

$$y(i) = r(i) + e(i)$$

rather than $r(i)$ directly. In this form we find that this model is a special case of our explicit EV statistical model in which only the dependent variable is subject to error:

$$\alpha_* x(i) = 0 \text{ where we observe } z_*(i) = x(i) + \delta_*(i) \text{ for } 1 \leq i \leq n,$$

with

$$\alpha_* = (\alpha_1, \alpha_2, \dots, \alpha_{p-1}, -1) ,$$

$$x(i) = (x_1(i), x_2(i), \dots, x_{p-1}(i), r(i)), x_1(i) = 1 ,$$

$$\delta_*(i) = (0, 0, \dots, 0, e(i)) ,$$

and

$$z_*(i) = (x_1(i), x_2(i), \dots, x_{p-1}(i), y(i)) = x(i) + \delta_*(i) .$$

Now taking account of the slight notational adjustments required, we may write down the normal equations for the GLS-estimators of this model from our results as

$$A_*^\dagger \alpha_*^T = 0$$

where

$$A_{\alpha_*}^\dagger =$$

$$\sum_{i=1}^n \sum_{j=1}^n \left(w_{ij}(\alpha_*) \left(\frac{(z_*^T(i) z_*(j) + z_*^T(j) z_*(i))}{2} \right)^\dagger - \mu_1(\alpha_*) \mu_j(\alpha_*) (\Sigma_{1j}^*)^\dagger \right),$$

$$W(\alpha_*) = [w_{ij}(\alpha_*)] = [\alpha_* \Sigma_{ij}^* \alpha_*^T]^{-1} = \Sigma_{d(\alpha_*)}^{-1},$$

$$\mu(\alpha_*) = (\mu_1(\alpha_*), \mu_2(\alpha_*), \dots, \mu_n(\alpha_*)) = \alpha_* Z_* W(\alpha_*),$$

and

$$Z_* = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_2(1) & x_2(2) & \dots & x_2(n) \\ x_3(1) & x_3(2) & \dots & x_3(n) \\ \vdots & \vdots & & \vdots \\ x_{p-1}(1) & x_{p-1}(2) & \dots & x_{p-1}(n) \\ y(1) & y(2) & \dots & y(n) \end{bmatrix} = (z_*^T(1), z_*^T(2), \dots, z_*^T(i)) .$$

We see immediately that

$$\Sigma_{ij}^* = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_{ij} \end{bmatrix} = E \delta_*^T(i) \delta_*(j)$$

where $\sigma_{ij} = Ee(i)e(j)$ and, therefore,

$$\sigma_{ij} = \alpha_* \sum_{ij}^* \alpha_*^T.$$

Consequently

$$W(\alpha_*) = [w_{ij}] = [\sigma_{ij}]^{-1} = \Sigma_e^{-1}$$

is independent of the unknown parameters $\alpha_1, \alpha_2, \dots$, and α_{p-1} . Moreover,

$$(\sum_{ij}^*)^\dagger = 0,$$

i.e., $(\sum_{ij}^*)^\dagger$ is the $(p-1) \times p$ -matrix whose elements are all 0, and so our system of normal equations reduces to

$$A_{\alpha_*}^\dagger \alpha_*^T = 0$$

where

$$A_{\alpha_*}^\dagger = \sum_{i=1}^n \sum_{j=1}^n w_{ij} \left(\frac{(z_*^T(i) z_*(j) + z_*^T(j) z_*(i))}{2} \right)^\dagger.$$

Now we note that whenever $A = A^T \in \Phi_{(n,n)}$ then

$$A = [a_{ij}] = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \left(\frac{F_{ij} + F_{ji}}{2} \right)$$

where $F_{ij} = f_i^T f_j$ and f_i is an n -dimensional vector with one as its i -th component and 0 elsewhere. The proof of this statement is direct:

$$\begin{aligned}
A &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} F_{ij} \\
&= \sum_{i=1}^n a_{ii} F_{ii} + \sum_{i>j} (a_{ij} F_{ij} + a_{ji} F_{ji}) \\
&= \sum_{i=1}^n a_{ii} F_{ii} + \sum_{i>j} a_{ij} (F_{ij} + F_{ji}) \\
&= \sum_{i=1}^n a_{ii} F_{ii} + 2 \sum_{i>j} a_{ij} \left(\frac{F_{ij} + F_{ji}}{2} \right) \\
&= \sum_{i=1}^n a_{ii} F_{ii} + \sum_{i>j} (a_{ij} + a_{ji}) \left(\frac{F_{ij} + F_{ji}}{2} \right) \\
&= \sum_{i=1}^n a_{ii} F_{ii} + \sum_{i>j} \left(a_{ij} \left(\frac{F_{ij} + F_{ji}}{2} \right) + a_{ji} \left(\frac{F_{ij} + F_{ji}}{2} \right) \right) \\
&= \sum_{i=1}^n a_{ii} \left(\frac{F_{ii} + F_{ii}}{2} \right) + \sum_{i \neq j} a_{ij} \left(\frac{F_{ij} + F_{ji}}{2} \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \left(\frac{F_{ij} + F_{ji}}{2} \right) .
\end{aligned}$$

Since $W = \Sigma_e^{-1}$, and Σ_e as a covariance matrix is symmetric, $W = [w_{ij}]$

is also a symmetric matrix, and from the result just stated we find

that

$$\begin{aligned}
A_{\alpha_*} &= \sum_{i=1}^n \sum_{j=1}^n w_{ij} \left(\frac{z_*^T(i) z_*(j) + z_*^T(j) z_*(i)}{2} \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n w_{ij} \left(\frac{z_*^T F_{ij} z_* + z_*^T F_{ji} z_*}{2} \right) \\
&= z_*^T \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij} \left(\frac{F_{ij} + F_{ji}}{2} \right) \right) z_* \\
&= z_*^T W z_*
\end{aligned}$$

Hence our normal equations can be written in the form

$$A_{\alpha_*}^\dagger \alpha_*^T = 0$$

where

$$A_{\alpha_*}^\dagger = (z_*^T W z_*)^\dagger = z_*^\dagger W z_*^T.$$

Now

$$z_*^\dagger = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1(1) & x_1(2) & \dots & x_1(n) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{p-1}(1) & x_{p-1}(2) & \dots & x_{p-1}(n) \end{bmatrix} = (x^T(1), \dots, x^T(n)) = X$$

and

$$z_*^T = [X^T, y^T] \quad \text{where} \quad y = (y(1), y(2), \dots, y(n)).$$

Consequently,

$$A_{\alpha_*}^{\dagger} = Z_*^{\dagger} W Z_*^T = XW [X^T, y^T] = [XWX^T, XWy^T]$$

and our equation system

$$A_{\alpha_*}^{\dagger} \alpha_*^T = 0$$

becomes

$$[XWX^T, XWy^T] \begin{bmatrix} \alpha^T \\ -1 \end{bmatrix} = 0, \quad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_{p-1}) ;$$

or, equivalently,

$$XWX^T \alpha^T = XWy^T ,$$

which is just the normal equation system developed first by Aitken and then by Plackett defining the generalized Gauss-Markoff least squares estimators.

Kummel-Lindley EV Model Specification

This model examined by Kummel in [89] and by Lindley in [91] is essentially as follows. They consider a single linear relationship between two variables,

$$s(i) = a + br(i) ,$$

where, for $1 \leq i \leq n$, the $s(i)$ and $r(i)$ are non-observable. Instead one observes

$$y(i) = s(i) + \xi(i) \quad \text{and} \quad x(i) = r(i) + \eta(i)$$

where the errors of observation $\delta(i) = (\eta(i), \xi(i))$ are assumed to be uncorrelated and identically distributed about $(0,0)$ with covariance matrix given by

$$\Sigma_{\delta} = \begin{bmatrix} \sigma_{\eta}^2 & 0 \\ 0 & \sigma_{\xi}^2 \end{bmatrix}$$

Further they assume that

$$\Sigma_{\delta} = \sigma_{\xi}^2 \begin{bmatrix} \mu & 0 \\ 0 & 1 \end{bmatrix}, \quad \mu = \sigma_{\eta}^2 \sigma_{\xi}^{-2},$$

is known up to the scalar multiple σ_{ξ}^2 . In this case, writing $\alpha = (a, b, -1)$, we find that

$$d(\alpha) = (a, b, -1) \begin{bmatrix} 1 & 1 & \dots & 1 \\ x(1) & x(2) & \dots & x(n) \\ y(1) & y(2) & \dots & y(n) \end{bmatrix} \begin{bmatrix} a \\ b \\ -1 \end{bmatrix} = (d_1(\alpha), d_2(\alpha), \dots, d_n(\alpha))$$

and that

$$\sigma_{ii}(\alpha) = E d_i(\alpha) d_i(\alpha) = \alpha \Sigma_{ii} \alpha^T = \sigma_{\xi}^2 (\mu b^2 + 1)$$

where

$$\Sigma_{11} = \sigma_{\xi}^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

Hence

$$\Sigma_d = \sigma_{\xi}^2(\mu b^2 + 1) I_n \quad \text{and} \quad \Sigma_d^{-1} = \frac{1}{\sigma_{\xi}^2(\mu b^2 + 1)} I_n ;$$

in addition we find that

$$\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \Sigma_{ij} \mu_j(\alpha) = \frac{1}{\sigma_{\xi}^2(\mu b^2 + 1)} \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mu \phi & 0 \\ 0 & 0 & \phi \end{bmatrix}$$

where

$$\phi = \phi(\alpha) = \frac{\alpha X X^T \alpha^T}{1 + \mu b^2} \quad \text{and} \quad \mu(\alpha) = \alpha X \Sigma_d^{-1} .$$

Consequently, under the Kummel-Lindley EV model specification, we obtain the normal equation system given by

$$\begin{bmatrix} n & \sum_{i=1}^n x(i) & \sum_{i=1}^n y(i) \\ \sum_{i=1}^n x(i) & \sum_{i=1}^n x^2(i) - \phi \mu & \sum_{i=1}^n x(i)y(i) \\ \sum_{i=1}^n y(i) & \sum_{i=1}^n x(i)y(i) & \sum_{i=1}^n y^2(i) - \phi \end{bmatrix} \begin{bmatrix} a \\ b \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Now $\phi(\alpha) = \sigma^2 Q(\alpha)$ where $Q(\alpha)$ is our GLS-likelihood function corresponding to the Kummel-Lindley EV model specification; and, consequently, if it were possible to specify the value ϕ_* of $\phi(\alpha)$ from an independent source either at the minimum value Q_0 of $Q(\alpha)$ or at the value $\hat{Q}_0$ of $Q(\alpha)$ evaluated at the GLS-estimators $\hat{a}$ and $\hat{b}$, then these estimators could be readily obtained as solution to the linear system of equations obtained by substitution of ϕ_* for ϕ in the non-linear system of GLS-normal equations given above. We would then find that

$$\hat{a} = \bar{y} - \hat{b} \bar{x}$$

and that

$$\hat{b} = \frac{\sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})}{\sum_{i=1}^n (x(i) - \bar{x})^2 - \phi_* \mu}.$$

Unfortunately such additional information is generally lacking and therefore we must give full consideration to the fact that ϕ depends on a and b too and that we are really dealing with a non-linear system of equations in b . Thus we find

$$\hat{\phi} = \frac{\sum_{i=1}^n (\hat{a} + \hat{b} x(i) - y(i))^2}{1 + \mu \hat{b}^2}$$

$$= \frac{\sum_{i=1}^n ((x(i) - \bar{x})\hat{b} - (y(i) - \bar{y}))^2}{1 + \mu \hat{b}^2} ;$$

and upon substituting this estimate of ϕ in the equation

$$\hat{b} \left(\sum_{i=1}^n (x(i) - \bar{x})^2 - \phi \mu \right) = \sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})$$

and rearranging terms while denoting deviations about the mean by

$\hat{z}(i) = z(i) - \bar{z}$, we obtain the non-linear equation

$$\mu \hat{b}^2 \sum_{i=1}^n \hat{x}(i) \hat{y}(i) + \hat{b} \sum_{i=1}^n (\hat{x}^2(i) - \mu \hat{y}^2(i)) - \sum_{i=1}^n \hat{x}(i) \hat{y}(i) = 0 ;$$

or, equivalently setting $\lambda = \mu^{-1}$, the non-linear equation

$$\hat{b}^2 \sum_{i=1}^n \hat{x}(i) \hat{y}(i) + \hat{b} \sum_{i=1}^n (\lambda \hat{x}^2(i) - \hat{y}^2(i)) - \sum_{i=1}^n \lambda \hat{x}(i) \hat{y}(i) = 0$$

which is the quadratic equation for the GLS-estimator $\hat{b}$ of the direction parameter b of the linear structural relationship

$$s(i) = a + b r(i)$$

obtained by Kummel and Lindley separately under their particular specification for the EV model.

Koopmans-Tintner EV Model Specification

We consider an implicit EV model

$$\alpha_1 r_1(i) + \alpha_2 r_2(i) + \dots + \alpha_p r_p(i) = 0$$

where the $r_k(i)$ are observed with errors $\delta_k(i)$:

$$x_k(i) = r_k(i) + \delta_k(i) ;$$

let

$$\delta(i) = (\delta_1(i), \delta_2(i), \dots, \delta_p(i))$$

for $1 \leq i \leq n$ and suppose that $E\delta(i) = 0$ and

$$\sum_{ij} = \delta_{ij} \Sigma = E\delta^T(i)\delta(j)$$

where the scalar δ_{ij} is the Kronecker delta (and is one for $i = j$ and zero otherwise) and Σ is a positive definite, symmetric matrix. For this model which represents the Koopmans-Tintner specification we find that

$$\Sigma_d = [\alpha \sum_{ij} \alpha^T] = [\delta_{ij} \alpha \Sigma \alpha^T] = (\alpha \Sigma \alpha^T) I_n$$

and, therefore,

$$W = \frac{1}{\alpha \Sigma \alpha^T} I_n = \Sigma_d^{-1} .$$

In addition

$$\mu(\alpha) = \alpha X W = \left(\frac{1}{\alpha \Sigma \alpha^T} \right) \alpha X$$

and we see that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \Sigma_{ij} \mu_j(\alpha) &= \left(\sum_{i=1}^n \mu_i^2(\alpha) \right) \Sigma \\ &= \left(\sum_{i=1}^n \left(\frac{\alpha x^T(i)}{\alpha \Sigma \alpha^T} \right)^2 \right) \Sigma \\ &= \frac{\lambda(\alpha)}{\alpha \Sigma \alpha^T} \Sigma \end{aligned}$$

where

$$\lambda(\alpha) = \sum_{i=1}^n \frac{(\alpha x^T(i))^2}{\alpha \Sigma \alpha^T} = \frac{\alpha X X^T \alpha^T}{\alpha \Sigma \alpha^T} .$$

Thus our GLS-normal equation system under the Koopmans-Tintner specification becomes

$$(\lambda(\alpha) \Sigma - X X^T) \alpha^T = 0 ;$$

and we see that this is again a non-linear system of equations implicitly defining the GLS-estimator $\hat{\alpha}$ of α . If $\lambda = \lambda(\alpha)$ were to be specified in advance this system of equations would be a linear system of homogeneous equation in the "unknown" α and for a non-trivial solution $\hat{\alpha}$ of this system to exist we would need

$$|\lambda(\alpha) \Sigma - X X^T| = 0 .$$

Provided the matrices appearing in this determinantal equation are non-singular, this characteristic equation has p -distinct characteristic roots, say

$$\lambda_1, \lambda_2, \dots, \text{ and } \lambda_p .$$

Now, since

$$\lambda(\alpha) = Q(\alpha)$$

where $Q(\alpha)$ is the GLS-likelihood function corresponding to the Koopmans-Tintner EV model specification and is to be minimized, we can obtain an independent estimate $\hat{\lambda}$ of λ by taking

$$\hat{\lambda} = \min\{\lambda_1, \lambda_2, \dots, \lambda_p\} .$$

Thus $\hat{\lambda}$ is determined so that (1) the system of linear equations obtained by substituting $\hat{\lambda}$ for λ in our non-linear system of GLS-normal equations has a non-trivial solution $\hat{\alpha}$; and (2) so that $\hat{\lambda}$ satisfies the condition that $Q(\hat{\alpha})$ is to be a minimum. In this connection we note that since,

$$(\hat{\lambda} \Sigma - XX^T) \hat{\alpha}^T = 0 ,$$

it follows that,

$$\hat{\lambda} \Sigma \hat{\alpha}^T = XX^T \hat{\alpha}^T ;$$

and upon pre-multiplication of this last expression by $\hat{\alpha}$ we see that

$$\hat{\lambda} \hat{\alpha} \Sigma \hat{\alpha}^T = \hat{\alpha} XX^T \hat{\alpha}^T$$

or

$$\hat{\lambda} = \frac{\hat{\alpha} XX^T \hat{\alpha}^T}{\hat{\alpha} \Sigma \hat{\alpha}^T} = Q(\hat{\alpha}) .$$

This then is the solution obtained when $\Sigma_{ij} = \delta_{ij} I_p$ by Koopmans in [84] and by Tintner in [143] when $\Sigma_{ij} = \delta_{ij} \Sigma$ and thus the GLS-estimators for the Koopmans-Tintner EV model specifications are obtained as a special case of our results.

Villegas EV Model Specification

Generalizing the Koopmans-Tintner EV model specification, Villegas in [152] studied the Koopmans-Tintner EV model under the more general assumption that

$$\Sigma_{ij} = k_{ij} \Sigma$$

where k_{ij} are known scalars rather than

$$\Sigma_{ij} = \delta_{ij} \Sigma$$

where δ_{ij} is the Kronecker delta scalar. He was led to this particular representation for the covariance matrices of the errors through problems dealing with intra-block analyses of incomplete block experimental designs. Under the Villegas EV model specification we see that

$$\Sigma_d = [\alpha \Sigma_{ij} \alpha^T] = \alpha \Sigma \alpha^T [k_{ij}] = (\alpha \Sigma \alpha^T) K$$

where we assume that K is a non-singular, symmetric matrix. Therefore,

$$W = \frac{1}{\alpha \Sigma \alpha^T} K^{-1} = \Sigma_d^{-1} \quad \text{and} \quad X W X^T = \frac{1}{\alpha \Sigma \alpha^T} X K^{-1} X^T .$$

Moreover we find that

$$\mu(\alpha) = \alpha X W = \frac{1}{\alpha \Sigma \alpha^T} \alpha X K^{-1}$$

and consequently

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) \Sigma_{ij} \mu_j(\alpha) &= \left(\sum_{i=1}^n \sum_{j=1}^n \mu_i(\alpha) k_{ij} \mu_j(\alpha) \right) \Sigma \\ &= \frac{\lambda_*(\alpha)}{\alpha \Sigma \alpha^T} \Sigma \end{aligned}$$

where

$$\lambda_*(\alpha) = (\alpha \Sigma \alpha^T) (\mu(\alpha) K \mu^T(\alpha)) .$$

Thus under the Villegas specification our GLS-normal equations become

$$(\lambda_* \Sigma - X K^{-1} X^T) \alpha^T = 0 .$$

Again we note that

$$\lambda_*(\alpha) = \frac{\alpha X K^{-1} X^T \alpha^T}{\alpha \Sigma \alpha^T} = Q(\alpha)$$

where $Q(\alpha)$ is the GLS-likelihood function for this EV model specification. Accordingly, as with the Koopmans-Tintner specification, we find that we can obtain a non-trivial GLS-estimator $\hat{\alpha}$ of α simply by taking λ_* as the minimum characteristic root $\hat{\lambda}_*$ of the characteristic equation

$$|\lambda_* \Sigma - X K^{-1} X^T| = 0$$

and then proceeding to solve the linear system of equations

$$(\lambda_* \Sigma - X K^{-1} X^T) \hat{\alpha}^T = 0$$

to obtain an $\hat{\alpha}$. This is Villegas's result derived by him under the assumptions noted above.

Beginning with the classical Gauss-Markoff-Aitken specification for the EV model in which only the dependent variable is assumed to be in error and proceeding to successively more general specifications for the EV model through Villegas's results derived under the assumptions that all variables are in error and that the variance-covariance matrix of these disturbances taken jointly can be represented as the direct product $K \otimes \Sigma$ of two matrices, we have found that these various approaches to the estimation problem for the EV model can be obtained as special cases of the results presented in Chapters 4 and 5. We also note, moreover, that although these previous results have appeared in the literature over the span of a little over 1-1/2 centuries, even the recent specifications still are of a special and specific nature in comparison with the general EV model that we have studied.

5.5. SUMMARY OF RESEARCH

In this last section we present a summary of the new results developed in the course of research for this thesis. In broad outline, this research has led to new extensions of the Dwyer-Macphail matrix derivative calculus; basic relations between the Dwyer-Macphail matrix derivative and the **Frechet-Gateaux** matrix differential have been obtained; and these notions have been applied to problems of parameter

estimation in some important stochastic structures. More specifically the following results comprise the major contributions of this work.

A review of the ordinary (or J-type) matrix derivative and the Dwyer-Macphail (or K-type) matrix derivative has led to a duality type theorem between these two kinds of matrix derivatives. This duality theorem is a generalization of the three basic rules that had been given for deriving these two kinds of matrix derivatives, one from the other, for matrix functions given as sums of products of constant and variable matrices. These three rules in the calculus of Dwyer-Macphail matrix derivatives, which can be given a purely symbolic interpretation, have been rigorously established by an application of this more basic duality type theorem and certain properties of the Tr-operator.

Five types of higher order matrix derivatives have been defined and their properties studied. To this author's knowledge, none of these five higher order matrix derivatives had been previously introduced or investigated. Our research has led to results which allow all of these five types of higher order matrix derivatives to be derived for a large class of matrix functions by direct reapplication of the three simple rules which previously were known to be valid only for the first order matrix derivatives of the J or K types mentioned above.

One of these five types of higher order matrix derivatives, namely the Dwyer-Macphail matrix derivative, has provided a basis for developing a matrix function of a ~~matrix~~ variable in an infinite series expansion which for the first time explicitly features its matrix derivatives.

Thus the introduction of these new higher order Dwyer-Macphail matrix derivatives and their use in developing a matrix function of a matrix variable in a series expansion are two new extensions of the Dwyer-Macphail matrix derivative calculus which maintain the simplicity of the original development of the calculus of first order Dwyer-Macphail matrix derivatives.

The Dwyer-Macphail matrix derivatives have been shown to be basically related to Frechet and Gateaux matrix differentials. We find in particular that the abstract matrix polynomials which define these matrix differentials are completely and uniquely determined in terms of corresponding Dwyer-Macphail matrix derivatives; and the first order Dwyer-Macphail matrix derivative and Frechet-Gateaux matrix differential can be used as a basis for extending the relation

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

between the derivative and differential of a scalar function $y = f(x)$ of a scalar variable to a matrix function of a matrix variable.

Using this extended calculus of Dwyer-Macphail matrix derivatives, maximum likelihood estimators or generalized least squares estimators have been derived for the first time for the unknown parameters of a linear structure relationship having both error-in-variables and errors-in-equations (i.e., an EVE statistical model) and without restrictions on the covariance matrices of the unknown errors, except of course, that the disturbances form a second order random process (i.e., that

the variances and covariances of the variable and equation errors are finite); and the maximum likelihood estimators of the structural parameters of a linear structural relationship are shown to be identical to the generalized least square estimators of these parameters.

Our derivation of the in general non-linear system of normal equations defining the generalized least squares estimators of the parameters of a linear structural relationship having variable and equation disturbances involves a new compound function differentiation theorem for matrix functions of matrix variables which features the Dwyer-Macphail matrix derivatives of the functions defining the compound matrix function. In addition, by using the relationship that exists between Dwyer-Machpail matrix derivatives and Frechet and Gateaux matrix differentials, we have established not only the necessary conditions for critical points of this complicated GLS-functional but, as important, we also have established a sufficient condition, in terms of restrictions among the observations and the covariances of the errors, that a critical point of this functional be a local minimum.

Next we show that our results on estimating the parameters of a linear relationship having only errors-in-variables (i.e., an EV statistical model) are general enough to include as special cases previously known and valid maximum likelihood or weighted least squares estimators for such stochastic structures derived by other authors under simpler assumptions. In particular we find that the celebrated

Gauss-Markoff-Aitken least squares estimators developed under the assumption of uncorrelated or correlated errors in only the dependent variable can be obtained from our results. The works of Norton and Parel who worked with the linear regression model in which all variables are subject to errors is seen to be a special EV model, and their results can be written down immediately once this identification with our work is made. Next the Kummel-Lindley weighted least squares estimators for the parameters of a two variable linear relationships, derived under the specification that the errors-in-variables are uncorrelated and the ratio of their variances known, are seen to be a special case of these general results. Then the Koopmans-Tintner maximum likelihood estimators for the parameters of a linear structure relationship in several variables in which all the variables are observed with errors which are intercorrelated within a multivariate observation, but not serially correlated between observations, are obtained by adopting this specification in our analysis. As the last special case, we derive from our general results the recent work of C. Villegas published in 1961 in the Annals of Mathematical Statistics in which he assumes that errors in all variables of a linear structure relationships are present, but not errors-in-equations, and that there errors are not necessarily identically and independently normally distributed but rather that they may have intercorrelations and serial correlations of a nature whereby they jointly have a variance-covariance matrix which can be represented

as the direct product of two known positive definite matrices.

As previously indicated we have obtained the results on ML- or GLS-estimation of the unknown parameters of a linear structural relationship having errors-in-variables and/or errors-in-equations by means of analytical derivations using the extended calculus of Dwyer-Macphail matrix derivatives. This is perhaps not the only derivation that can be given, nor necessarily the shortest approach. It is well known, of course, that the ordinary least squares estimators for the parameters of a linear structural relationship in which only the dependent variable is subject to error can be characterized geometrically as an orthogonal projection on a linear subspace.

In the more general EV or EVE model when all variables are non-observable and the covariance matrix of the errors is assumed to be a known symmetric, positive definite matrix, but not necessarily diagonal, etc., the underlying geometry of the ML- or GLS-estimation procedure may also be of such a nature that one is lead directly to the normal equation systems, obtained in this work by analysis, through the application of certain linear transformations and a geometrical interpretation of their significance. Presently, however, it appears that the geometry of the ML- or GLS-estimators for the general EV and EVE statistical models is complicated and that the problem is not entirely simplified, for example, by first applying a linear transformation to the covariance operator in order to reduce it to an identity

matrix, thus working with uncorrelated rather than correlated observations, since such a transformation leads to a more involved system of restraints imposed by the underlying linear structural relationship than appears in the untransformed problem.

Finally, we point out that the full contribution of our findings in this area of statistics stems not only from the several new results on this estimation problem which have been obtained, but from the following observations as well: First, the application of the Dwyer-Macphail matrix derivative calculus in this work clearly demonstrates the potential use of these same techniques for deriving the non-linear normal systems for the ML- or GLS-estimators of the parameters of a multiple equation EVE model whose variable and equation errors are both intercorrelated and serially correlated. Second, either by direct analysis of our non-linear system of normal equations for the general single equation EVE statistical model or by Monte Carlo techniques, the ML- or GLS-estimators can be expressed as functions of the variances and covariances of the disturbances, thus permitting the effects of variable or equation errors on these estimators to be evaluated more fully in a particular application than was previously possible.

BIBLIOGRAPHY

1. Acton, F. S., Analysis of Straight-Line Data, John Wiley and Sons, Inc., New York, 1959.
2. Aitken, A. C., "On least squares and linear combinations of observations," Proc. Roy. Soc. of Edinburgh, 55 (1934), 42-48.
3. Alexiewicz, A., and Orlicz, W., "On the differentials in Banach spaces," Ann. Soc. Polon. Math., 25 (1952), 95-99.
4. Allen, R.G.D., "The assumptions of linear regression," Economica, New Series, 6.(1939), 191-204.
5. Anderson, R. L., "The problem of autocorrelation in regression analysis," Jour. Amer. Stat. Assoc., 44 (1949), 113-129.
6. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., New York, 1958.
7. Anderson, T. W., and Rubin, H., "Estimation of the parameters of a single equation in a complete system of stochastic equations," Ann. Math. Stat., 20 (1949), 46-63.
8. Anderson, T. W., and Rubin, H., "The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations," Ann. Math. Stat., 21 (1950), 570-582.
9. Austen, A.E.W., and Pelzer, H., "Linear curves of best fit," Nature, 157 (1946), 693-694.
10. Bartlett, M. S., "Fitting a straight line when both variables are subject to error," Biometrics, 5 (1949), 207-212.
11. Barton, D. E., and David, F. N., "Models of functional relationships illustrated on astronomical data," Bull. Inst. Int. Stat., 37 (1960), 9-33.
12. Basmann, R. L., "A generalized classical method of linear estimation of coefficients in structural equations," Econometrica, 25 (1957), 77-83.
13. Basmann, R. L., "An expository note on estimation of simultaneous structural equations," Biometrics, 16 (1960), 464-480.

14. Beech, D. G., "Some notes on the precision of the gradient of an estimated straight line," Appl. Stat., 10 (1961), 14-31.
15. Bentzel, R., and Wold, H., "On statistical demand analysis from the viewpoint of simultaneous equations," Skandinavisk Aktuarietidskrift, 29 (1946), 95-114.
16. Berkson, J., "Are there two regressions?" Jour. Amer. Stat. Assoc., 45 (1950), 164-180.
17. Berkson, J., "Estimation by least squares and by maximum likelihood," Proc. Third Berkeley Symposium, 1 (1956), 1-11.
18. Box, G.E.P., "Fitting empirical data," Ann. N. Y. Acad. Sci., 86 (1960), 792-816.
19. Brown, D. C., A Treatment of Analytical Photogrammetry With Emphasis on Ballistic Camera Applications, RCA Data Reduction Technical Report, No. 39, 1957; in particular Appendix A, "A treatment of the general problem of least squares."
20. Cherubino, S., Calcolo delle Matrici, Consiglio Nazionale delle Ricerche Monografie Matematiche 4, Edizioni Cremonese, Rome, 1957; in particular Chapter 4, 243-290.
21. Chipman, J. S., and Rao, M. M., On the Use of Idempotent Matrices in the Treatment of Linear Restrictions in Regression Analysis, Department of Statistics, Univ. of Minn., Tech. Rep. No. 10, 1959.
22. Cochran, D., and Orcutt, G. H., "Application of least-squares regression to relationships containing autocorrelated error terms," Jour. Amer. Stat. Assoc., 44 (1949), 32-61.
23. Coy, J. W., A Differential Calculus for Functions of Matrices, Ph.D. Dissertation, Univ. of Mich., 1955.
24. Cramér, H., Mathematical Methods of Statistics, Princeton Univ. Press, Princeton, N. J., 1946.
25. Creasy, M. A., "Confidence limits for the gradient in the linear functional relationship," Jour. Roy. Stat. Soc., Series B, 18 (1956), 65-69.
26. David, F. N., and Neyman, J., "Extension of the Markoff theorem on least squares," Stat. Res. Mem., 2 (1938), 105.

27. Deming, W. E., Statistical Adjustment of Data, John Wiley and Sons, Inc., New York, 1943.
28. Drion, E. F., "Estimation of the parameters of a straight line and of the variances of the variables if they are both subject to error," Indagationes Mathematicae, Proceeding of the Section of Science, Series A, 54 (1951), 256-260.
29. Durbin, J., "Errors in variables," Rev. Inst. Int. Statistique, 22 (1954), 23-32.
30. Durbin, J., and Watson, G. S., "Testing for serial correlation in least-squares regression," I and II, Biometrika, 37 (1950), 409-428; and 38 (1951), 159-178.
31. Dwyer, P. S., "Generalizations of a gaussian theorem," Ann. Math. Stat., 29 (1958), 106-117.
32. Dwyer, P. S., and Macphail, M. S., "Symbolic matrix derivatives," Ann. Math. Stat., 19 (1948), 517-534.
33. Eisenhart, C., "Interpretation of certain regression methods and their use in biological and industrial research," Ann. Math. Stat., 10 (1939), 162-186.
34. Ferguson, T. S., "A method of generating best asymptotically normal estimates," Ann. Math. Stat., 29 (1958), 1046-1062.
35. Frazer, R. A., Duncan, W. J., and Collar, A. R., Elementary Matrices and Some Applications to Dynamics and Differential Equations, Cambridge Univ. Press, 1936.
36. Frechet, M., "Une definition fonctionnelle des polynomes," Nouv. Ann. Math., (4) 9 (1909), 145-162.
37. Frechet, M., "Sur les fonctionnelles continues," Ann. Ecole Norm. Sup., (3) 27 (1910), 193-216.
38. Frechet, M., "La notion de differentielle dans l'analyse generale," Ann. Ecole Norm. Sup., (3) 42 (1925), 293-323.
39. Gateaux, R., "Fonctions d'une infinite des variables independantes," Bull. Soc. Math. de France, 47 (1919), 70-96.
40. Gateaux, R., "Sur les fonctionnelles continues et les fonctionnelles analytiques," Bull. Soc. Math. de France, 50 (1922), 1-21.

41. Gauss, C. F., Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrich Gauss, Berlin, 1887. Translation from Latin by A. Borsch and P. Simon.
42. Geary, R. C., "Inherent relations between random variables," Proc. Roy. Irish Acad., 47 (1942), 63-76.
43. Geary, R. C., "Relations between statistics: the general and the sampling problem," Proc. Roy. Irish Acad., 49 (1943), 177-196.
44. Geary, R. C., "Studies in relations between economic time series," Jour. Roy. Stat. Soc., Series B, 10 (1948), 140.
45. Geary, R. C., "Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown," Econometrica, 17 (1949), 30-58.
46. Gibson, W. M., and Jowett, G. H., "Three-group regression analysis. Part I. Single regression analysis," Appl. Stat., 6 (1957), 114-122.
47. Gini, C., "Sull interpolazione di una retta quando i valore della variabile indipendente sono affetti da errori accidentali," Metron, 1 (1921), 63-82.
48. Girshick, M. A., and Haavelmo, T., "Statistical analysis of the demand for food: examples of simultaneous estimation of structural equations," Econometrica, 15 (1947), 79-110.
49. Graves, L. M., "Riemann integration and Taylor's theorem in general analysis," Trans. Amer. Math. Soc., 29 (1927), 163-177.
50. Graves, L. M., "Topics in the functional calculus," Bull. Amer. Math. Soc., 41 (1935), 641-662.
51. Graves, L. M., and Hildebrandt, T. H., "Implicit functions and their differentials in general analysis," Trans. Amer. Math. Soc., 29 (1927), 127-153.
52. Grenander, U., "On the estimation of regression coefficients in the case of an autocorrelated disturbance," Ann. Math. Stat., 25 (1954), 252-272.
53. Haavelmo, T., "The statistical implications of a system of simultaneous equations," Econometrica, 11 (1943), 1-12.

54. Haavelmo, T., "The probability approach in econometrics," Econometrica, Sup., 12 (1944).
55. Halmos, P. R., Finite Dimensional Vector Spaces, Annals of Mathematics Studies No. 7, Princeton Univ. Press, Princeton, New Jersey, 1953.
56. Hartley, H. O., "The estimation of non-linear parameters by 'internal least squares,'" Biometrika, 35 (1948), 32-45.
57. Hausdorff, F., "Zur theorie der systeme zahlen," Mathematisch and Physische Klasse, 52 (1900), 43-61.
58. Hemelrijk, J., "Construction of a confidence region for a line," Proc. Ned. Akad. Wetensch., 52 (1949), 995-1005.
59. Highberg, I. E., "A note on abstract polynomials in complex spaces," J. de Math., (9) 16 (1937), 307-314.
60. Hille, E., and Phillips, R. S., Functional Analysis and Semi-Groups, American Mathematical Society Colloquium Pub., Waverly Press, Inc., Baltimore, Md., 31, 1948; rev. ed., 1957; in particular, Chapters 3 and 26.
61. Hood, W. C., and Koopmans, T. C., Studies in Econometric Method, Cowles Comm. Monograph 14, John Wiley and Sons, Inc., 1953.
62. Hotelling, H., "Relations between two sets of variates," Biometrika, 28 (1926), 321-377.
63. Hotelling, H., "Analysis of a complex of statistical variables into principal components," Jour. of Educ. Psych., 24 (1933), 417-498.
64. Hotelling, H., "The selection of variates for use in prediction with some comments on the general problem of nuisance parameters," Ann. Math. Stat., 11 (1940), 271.
65. Housner, G. W., and Brennan, J. F., "The estimation of linear trends," Ann. Math. Stat., 19 (1948), 380-388.
66. Hurwicz, L., "Stochastic models of economic fluctuations," Econometrica, 12 (1944), 114-124.
67. Hyers, D. H., "A generalization of Frechet's differential," Proc. Nat. Acad. Sci., U.S.A., 27 (1941), 315-316.

68. Jeeves, T. A., "Identification and estimation of linear manifolds in n-dimensions," Ann. Stat., 25 (1954), 714-723.
69. Jessop, W. N., A Critical Review of Work Relating to One-Line Representations of the Relationship Between Variables, U. S. Steel Co., Res. and Dev. Dept., Stat. Sect., R. 001, 1950.
70. Jessop, W. N., "One line or two?" Appl. Stat., 2 (1952), 131-137.
71. Johnston, J., Econometric Methods, McGraw-Hill, Inc., New York, 1953.
72. Jones, H. E., "Some geometrical considerations in the general theory of fitting lines and planes," Metron, 13 (1937), 21.
73. Jowett, G. H., "Least squares regression analysis for trend-reduced time series," Jour. Roy. Stat. Soc., 17 (1955), 91-104.
74. Jows, H., "Linear regression functions with neglected variables," Jour. Amer. Stat. Assoc., 41 (1946), 356-369.
75. Kantorovich, L. V., "Functional analysis and applied mathematics," Uspekhi Mat. Nauk., 3 (1948), 89-185; translation by C. D. Benster, National Bureau of Standards Report 1509, G. E. Forsythe, ed., 1952.
76. Kantorovich, L. V., and Akiloff, G. P., Functional Analysis in Normed Spaces, Russian, 1959; in particular Chapter 17, 589-621.
77. Kendall, M. G., "Regression, structure and functional relationship," I and II, Biometrika, 38 (1951), 11-25; 39 (1952), 96-108.
78. Kendall, M. G., A Course in Multivariate Analysis, Griffin, London, 1957; in particular Chapter 2.
79. Kendall, M. G., and Stuart, A., The Advanced Theory of Statistics I. Distribution Theory, Griffin, London, 1958.
80. Kendall, M. G., and Stuart, A., The Advanced Theory of Statistics II. Inference and Relationships, Griffin, London, 1961; in particular Chapter 29, 375-418.
81. Kendall, M. G., and Stuart, A., The Advanced Theory of Statistics III. Planning and Analysis and Time Series, Griffin, London, unpublished.

82. Kiefer, J., and Wolfowitz, J., "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," Ann. Math. Stat., 27 (1956), 887-906.
83. Konijn, H. S., "Identification and estimation in a simultaneous equations model with errors in the variables," Econometrica, 30 (1962), 79-87.
84. Koopmans, T. C., Linear Regression Analysis of Economic Time Series, DeErven F. Bohn, Haarlem, Netherlands, 1937.
85. Koopmans, T. C., "Statistical estimation of simultaneous economic relationships," Jour. Amer. Stat. Assoc., 40 (1945), 448-466.
86. Koopmans, T. C., "Identification problems in economic model construction," Econometrica, 17 (1949), 125-144.
87. Koopmans, T. C., Ed., Statistical Inference in Dynamic Economic Models, Cowles Comm. Monograph 10, John Wiley and Sons, Inc., 1950.
88. Koopmans, T. C., and Reiersøl, O., "The identification of structural characteristics," Ann. Math. Stat., 21 (1950), 165-181.
89. Kummel, C. H., "Reduction of observed equations which contain more than one observed quantity," The Analyst, 6 (1879), 97-105.
90. **Kruskal**, W., "The coordinate-free approach to gauss-Markov estimation and its applications to missing and extra observations," Proceedings of the Fourth Berkeley Symposium, 4 (1961), 435-451.
91. Lindley, D. V., "Regression lines and the linear functional relationship," Jour. Roy. Stat. Soc., Series B, Supp. 9 (1947), 219-244.
92. Lindley, D. V., "Estimation of a functional relationship," Biometrika, 40 (1953), 47-49.
93. Linnik, Y. V., Methods of Least Squares and Principles of the Theory of Observations, Russian, 1958. Translated, Pergamon Press, 1961.
94. Lorch, E. R., "The theory of analytic functions in normed abelian vector spaces," Trans. Amer. Math. Soc., 54 (1943), 414-425.
95. Lusternik, L. A., and Sobolev, V. J., Elements of Functional Analysis, Frederick Ungar Pub. Co., New York, 1961.
96. Mac Duffee, C. C., The Theory of Matrices, J. Springer Co., Berlin, 1933.

97. Mac Duffee, C. C., Vectors and Matrices, Mathematical Association of America, Menasha, Wisc., 1943.
98. Mac Duffee, C. C., The Theory of Matrices, Chelsea Pub. Co., New York, 1946.
99. Madansky, A., "The fitting of straight lines when both variables are subject to error," Jour. Amer. Stat. Assoc., 54 (1959), 173-205.
100. Mann, H. B., and Wald, A., "On the statistical treatment of linear stochastic difference equations," Econometrica, 11 (1943), 173-220.
101. Markoff, A. A., Wahrscheinlichkeitsrechnung, Leipzig, 1912.
102. Martin, R. S., and Michal, A. D., "Some expansions in vector spaces," J. de Math., (9) 13 (1934), 69-91.
103. Michal, A. D., "General differential geometries and related topics," Bull. Amer. Math. Soc., 45 (1939), 529-563.
104. Michal, A. D., "On bounds of polynomials in hyperspheres and Frechet-Michal derivatives in real and complex normed linear spaces," Math. Mag., 27 (1954), 119-126.
105. Moran, P.A.P., "A test of significance for an unidentifiable relation," Jour. Roy. Stat. Soc., Series B, 18 (1956), 61-64.
106. Nair, K. R., and Banerjee, K. S., "A note on fitting straight lines if both variables are in error," Sankhya, 6 (1941-42), 331.
107. Nair, K. R., and Sheivastava, M. P., "On a simple method of curve fitting," Sankhya, 6 (1942-43), 121-132.
108. Nerlove, M., "Effects of serial correlation in the error terms," in R. J. Foote, Analytical Tools for Studying Demand and Price Structure, U. S. Department of Agriculture, Wash., D.C., 1958.
109. Neyman, J., "Remarks on a paper by E. C. Rhodes," Jour. Roy. Stat. Soc., 100 (1937), 50-57.
110. Neyman, J., "Existence of consistent estimate of the directional parameter in a linear structural relationship between two variables," Ann. Math. Stat., 22 (1951), 497-512.

111. Neyman, J., and Scott, E. L., "Consistent estimates based on partially consistent observations," Econometrica, 16 (1948), 1-32.
112. Neyman, J., and Scott, E. L., "On certain methods of estimating the linear structural relation," Ann. Math. Stat., 22 (1951), 352-361; correction 23 (1952), 115.
113. Norton, K. A., Fitting an Arbitrary Functional Relationship by Least Squares with All of the Variables Subject to Error, U. S. Department of Commerce, National Bureau of Standards, Report No. 5068, 1957.
114. Parel, C. P., A Matrix Derivation of Generalized Least Squares Linear Regression With All Variables Subject to Error, Ph.D. Dissertation, Univ. of Mich., 1958.
115. Pearson, K., "On lines and planes of closest fit to systems of points in space," Phil. Mag., Series 4, 2 (1901), 559-572.
116. Pettis, B. J., "Differentiation in Banach spaces," Duke Math. Jour., 5 (1939), 254-269.
117. Pietra, G., "Interpolating plane curves," Metron, 3 (1924), 311.
118. Plackett, R. L., "A historical note on the method of least squares," Biometrika, 36 (1949), 458-460.
119. Reiersøl, O., "Confluence analysis by means of lag moments and other methods of confluence analysis," Econometrica, 9 (1941), 1-24.
120. Reiersøl, O., "Confluence analysis by means of instrumental sets of variables," Arkiv for Matematik, Astron. och Fysik, 32 (1945), 1-119.
121. Reiersøl, O., "Identifiability of a linear relation between variables which are subject to error," Econometrica, 18 (1950), 375-389.
122. Rinehart, R. F., "The equivalence of definitions of a matrix function," Amer. Math. Monthly, 62 (1955), 395-414.
123. Rinehart, R. F., "Extension of the derivative concept for functions of matrices," Proc. Amer. Math. Soc., 8 (1957), 329-335.
124. Rhodes, E. C., "On lines and planes of closest fit," Phil. Mag., Series 7, 3 (1927), 357-364.

125. Roos, C. F., "A general variant criterion of fit for lines and planes when all variates are subject to error," Metron, 13 (1937), 3.
126. Sargan, J. D., "The maximum likelihood estimation of economic relationships with autoregressive residuals," Econometrica, 29 (1961), 414-426.
127. Savage, L. J., The Foundation of Statistics, John Wiley and Sons, Inc., New York, 1951.
128. Scott, E. L., "Note on consistent estimates on the linear structural relationship between two variables," Ann. Math. Stat., 21 (1950), 284-288.
129. Seares, F. H., "Regression lines and the functional relationship," Astrophys. Jour., 100 (1944), 255-263.
130. Siddiqui, M., "Covariances of least squares estimates when residuals are correlated," Ann. Math. Stat., 29 (1958), 1251-1256.
131. Smith, H. F., Estimating a Linear Functional Relation, ASTIA Document, No. 85543.
132. Stuart, A., "The correlation between variate-values and ranks in samples from a continuous distribution," Brit. Jour. Stat. Psych., 7 (1954), 37-44.
133. Suits, D. B., "Forecasting and analysis with econometric models," Jour. Amer. Econ. Assoc., 52 (1962), 104-132.
134. Taylor, A. E., "Analysis in complex Banach spaces," Bull. Amer. Math. Soc., 49 (1943), 652-669.
135. Taylor, A. E., Introduction to Functional Analysis, John Wiley and Sons, Inc., New York, 1958.
136. Taylor, W. F., "Distance functions and regular best asymptotically normal estimates," Ann. Math. Stat., 24 (1953), 85-92.
137. Theil, H., "A rank-invariant method of linear and polynomial regression analysis," Indagationes Math., 12 (1950), 85-91.
138. Theil, H., Economic Forecasts and Policy, North-Holland Pub. Co., Amsterdam, 1958.
139. Theil, H., and Kloeck, T., "The statistics of systems of simultaneous economic relationships," Bull. Inst. Int. Statistique, 37 (1960), 345-368.

140. Theil, H., and van Yzeren, J., "On the efficiency of Wald's method of fitting straight lines," Rev. Inst. Int. Statistique, 24 (1956), 17-26.
141. Thompson, G. H., "A hierarchy without a general factor," Brit. Jour. Psych., 8 (1916), 271-281.
142. Tintner, G., "An application of the variate difference method to multiple regression," Econometrica, 12 (1944), 97-113.
143. Tintner, G., "Multiple regression for systems of equations," Econometrica, 14 (1946), 5-36.
144. Tintner, G., Econometrics, John Wiley and Sons, Inc., New York, 1952.
145. Tintner, G., "A note on economic aspects of the theory of errors in time series," Quart. Jour. Econ., 53 (1958), 141.
146. Todd, J., Ed., Survey of Numerical Analysis, McGraw-Hill Book Co., Inc., New York, 1962; in particular Chapter 16, 485-491.
147. Tukey, J. W., "Components in regression," Biometrics, 7 (1951), 33-69.
148. Tysver, J. B., Inherent Errors in Matrices with Statistical Applications, Ph.D Dissertation, Univ. of Mich., 1958.
149. Thrall, R. M., and Tornheim, L., Vector Spaces and Matrices, John Wiley and Sons, Inc., New York, 1957.
150. van Uven, M. J., "Adjustment of N points (in n-dimensional space) to the best linear (n-1)-dimensional space," Koninklijke Akademie van Wetenschappen te Amsterdam, Proc. Sect. Sci., 23 (1930), 143-307.
151. van Uven, M. J., Mathematical Treatment of the Results of Agricultural and Other Experiments, P. Noordhoff N. V.-Groningen Batavin, Netherlands, 1935.
152. Villegas, C., "Maximum likelihood estimation of a linear functional relationship," Ann. Math. Stat., 32 (1961), 1048-1062.
153. Wald, A., "The fitting of straight lines if both variables are subject to error," Ann. Math. Stat., 11 (1940), 284-300.

154. Wald, A., "Tests of statistical hypotheses concerning several parameters when the number of observations is large," Amer. Math. Soc. Trans., 54 (1943), 426-482.
155. Watson, G. S., Serial Correlation in Regression Analysis, Ph.D. Dissertation, University of North Carolina, 1951.
156. Whittaker, E. T., and Robinson, G., The Calculus of Observations, Blackie, London, 4th edition, 1944.
157. Winsor, C. P., "Which regression?" Biometrics Bull., 2 (1946), 101-109.
158. Wise, J., "Regression analysis of relationships between autocorrelated time series," Jour. Roy. Stat. Soc., Series B, 18 (1956), 240-256.
159. Wold, H., "On least squares regression with autocorrelated variables and residuals," Bull. Inst. Int. Stat., 32 (1950), 277-289.
160. Wold, H., and Jureen, L., Demand Analysis: A Study in Econometrics, John Wiley and Sons, Inc., 1953.
161. Wold, H., "Causal inference from observation data: A review of ends and means," Jour. Roy. Stat. Soc., Series A, 119 (1956), 28-61.
162. Wolfowitz, J., "Consistent estimators of the parameters of a linear structural relationship," Skandinavisk Aktuarietidskrift, 32 (1952), 132-151.
163. Wolfowitz, J., "Estimation by the minimum distance method," Ann. Inst. Stat. Math., 5 (1953), 9-23.
164. Wolfowitz, J., "Estimation of the components of stochastic structures," Proc. Nat. Acad. Sci., U.S.A., 40 (1954), 602-606.
165. Wolfowitz, J., "The minimum distance method," Ann. Math. Stat., 28 (1957), 89-110.
166. Zellner, A., "Econometric estimation with temporally dependent disturbance terms," Int. Econ. Rev., 2 (1961), 164-178.
167. Zorn, M. A., "Gateaux differentiability and essential boundedness," Duke Math. Jour., 12 (1945), 579-583.

168. Zorn, M. A., "Derivatives and Frechet differentials," Bull. Amer. Math. Soc., 52 (1946), 133-137.
169. Zucker, L. M., "Evaluation of slope and intercept of straight lines," Human Biology, 19 (1947), 231-259.
170. Zuffi, C., Ed., Scritti Matematici in Onore di Filippo Sibirani, Bologna, 1957; in particular di Vincenzo Amato (a Catania), "Funzioni di matrice," 5-11.