Matrix Differential Calculus with Applications in Statistics and Econometrics

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter E. Shewhart and Samuel S. Wilks*

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, and Ruey S. Tsay*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane, and Jozef L. Teugels*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological, and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of the titles in this series can be found at
http://www.wiley.com/go/wsps

Matrix Differential Calculus with Applications in Statistics and Econometrics

Third Edition

Jan R. Magnus
Department of Econometrics and Operations Research,
Vrije Universiteit Amsterdam, The Netherlands

and

Heinz Neudecker†
Amsterdam School of Economics,
University of Amsterdam, The Netherlands

WILEY

This third edition first published 2019

Edition History
John Wiley & Sons (1e, 1988) and John Wiley & Sons (2e, 1999)

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data applied for
ISBN: 9781119541202

Cover design by Wiley
Cover image: © phochi/Shutterstock
Typeset by the author in LaTeX

10 9 8 7 6 5 4 3 2 1

Contents

Part Five — The linear model

Preface

Preface to the first edition

There has been a long-felt need for a book that gives a self-contained and unified treatment of matrix differential calculus, specifically written for econometricians and statisticians. The present book is meant to satisfy this need. It can serve as a textbook for advanced undergraduates and postgraduates in econometrics and as a reference book for practicing econometricians. Mathematical statisticians and psychometricians may also find something to their liking in the book.

When used as a textbook, it can provide a full-semester course. Reasonable proficiency in basic matrix theory is assumed, especially with the use of partitioned matrices. The basics of matrix algebra, as deemed necessary for a proper understanding of the main subject of the book, are summarized in Part One, the first of the book's six parts. The book also contains the essentials of multivariable calculus but geared to and often phrased in terms of differentials.

The sequence in which the chapters are being read is not of great consequence. It is fully conceivable that practitioners start with Part Three (Differentials: the practice) and, dependent on their predilections, carry on to Parts Five or Six, which deal with applications. Those who want a full understanding of the underlying theory should read the whole book, although even then they could go through the necessary matrix algebra only when the specific need arises.

Matrix differential calculus as presented in this book is based on differentials, and this sets the book apart from other books in this area. The approach via differentials is, in our opinion, superior to any other existing approach. Our principal idea is that differentials are more congenial to multivariable functions as they crop up in econometrics, mathematical statistics, or psychometrics than derivatives, although from a theoretical point of view the two concepts are equivalent.

The book falls into six parts. Part One deals with matrix algebra. It lists, and also often proves, items like the Schur, Jordan, and singular-value decompositions; concepts like the Hadamard and Kronecker products; the vec operator; the commutation and duplication matrices; and the Moore-Penrose

inverse. Results on bordered matrices (and their determinants) and (linearly restricted) quadratic forms are also presented here.

Part Two, which forms the theoretical heart of the book, is entirely devoted to a thorough treatment of the theory of differentials, and presents the essentials of calculus but geared to and phrased in terms of differentials. First and second differentials are defined, 'identification' rules for Jacobian and Hessian matrices are given, and chain rules derived. A separate chapter on the theory of (constrained) optimization in terms of differentials concludes this part.

Part Three is the practical core of the book. It contains the rules for working with differentials, lists the differentials of important scalar, vector, and matrix functions (*inter alia* eigenvalues, eigenvectors, and the Moore-Penrose inverse) and supplies 'identification' tables for Jacobian and Hessian matrices.

Part Four, treating inequalities, owes its existence to our feeling that econometricians should be conversant with inequalities, such as the Cauchy-Schwarz and Minkowski inequalities (and extensions thereof), and that they should also master a powerful result like Poincaré's separation theorem. This part is to some extent also the case history of a disappointment. When we started writing this book we had the ambition to derive all inequalities by means of matrix differential calculus. After all, every inequality can be rephrased as the solution of an optimization problem. This proved to be an illusion, due to the fact that the Hessian matrix in most cases is singular at the optimum point.

Part Five is entirely devoted to applications of matrix differential calculus to the linear regression model. There is an exhaustive treatment of estimation problems related to the fixed part of the model under various assumptions concerning ranks and (other) constraints. Moreover, it contains topics relating to the stochastic part of the model, viz. estimation of the error variance and prediction of the error term. There is also a small section on sensitivity analysis. An introductory chapter deals with the necessary statistical preliminaries.

Part Six deals with maximum likelihood estimation, which is of course an ideal source for demonstrating the power of the propagated techniques. In the first of three chapters, several models are analysed, *inter alia* the multivariate normal distribution, the errors-in-variables model, and the nonlinear regression model. There is a discussion on how to deal with symmetry and positive definiteness, and special attention is given to the information matrix. The second chapter in this part deals with simultaneous equations under normality conditions. It investigates both identification and estimation problems, subject to various (non)linear constraints on the parameters. This part also discusses full-information maximum likelihood (FIML) and limited-information maximum likelihood (LIML), with special attention to the derivation of asymptotic variance matrices. The final chapter addresses itself to various psychometric problems, *inter alia* principal components, multimode component analysis, factor analysis, and canonical correlation.

All chapters contain many exercises. These are frequently meant to be complementary to the main text.

A large number of books and papers have been published on the theory and applications of matrix differential calculus. Without attempting to describe their relative virtues and particularities, the interested reader may wish to consult Dwyer and Macphail (1948), Bodewig (1959), Wilkinson (1965), Dwyer (1967), Neudecker (1967, 1969), Tracy and Dwyer (1969), Tracy and Singh (1972), McDonald and Swaminathan (1973), MacRae (1974), Balestra (1976), Bentler and Lee (1978), Henderson and Searle (1979), Wong and Wong (1979, 1980), Nel (1980), Rogers (1980), Wong (1980, 1985), Graham (1981), McCulloch (1982), Schönemann (1985), Magnus and Neudecker (1985), Pollock (1985), Don (1986), and Kollo (1991). The papers by Henderson and Searle (1979) and Nel (1980), and Rogers' (1980) book contain extensive bibliographies.

The two authors share the responsibility for Parts One, Three, Five, and Six, although any new results in Part One are due to Magnus. Parts Two and Four are due to Magnus, although Neudecker contributed some results to Part Four. Magnus is also responsible for the writing and organization of the final text.

We wish to thank our colleagues F. J. H. Don, R. D. H. Heijmans, D. S. G. Pollock, and R. Ramer for their critical remarks and contributions. The greatest obligation is owed to Sue Kirkbride at the London School of Economics who patiently and cheerfully typed and retyped the various versions of the book. Partial financial support was provided by the Netherlands Organization for the Advancement of Pure Research (Z. W. O.) and the Suntory Toyota International Centre for Economics and Related Disciplines at the London School of Economics.

London/Amsterdam — Jan R. Magnus
April 1987 — Heinz Neudecker

Preface to the first revised printing

Since this book first appeared — now almost four years ago — many of our colleagues, students, and other readers have pointed out typographical errors and have made suggestions for improving the text. We are particularly grateful to R. D. H. Heijmans, J. F. Kiviet, I. J. Steyn, and G. Trenkler. We owe the greatest debt to F. Gerrish, formerly of the School of Mathematics in the Polytechnic, Kingston-upon-Thames, who read Chapters 1–11 with awesome precision and care and made numerous insightful suggestions and constructive remarks. We hope that this printing will continue to trigger comments from our readers.

London/Tilburg/Amsterdam — Jan R. Magnus
February 1991 — Heinz Neudecker

Preface to the second edition

A further seven years have passed since our first revision in 1991. We are happy to see that our book is still being used by colleagues and students. In this revision we attempted to reach three goals. First, we made a serious attempt to keep the book up-to-date by adding many recent references and new exercises. Second, we made numerous small changes throughout the text, improving the clarity of exposition. Finally, we corrected a number of typographical and other errors.

The structure of the book and its philosophy are unchanged. Apart from a large number of small changes, there are two major changes. First, we interchanged Sections 12 and 13 of Chapter 1, since complex numbers need to be discussed before eigenvalues and eigenvectors, and we corrected an error in Theorem 1.7. Second, in Chapter 17 on psychometrics, we rewrote Sections 8–10 relating to the Eckart-Young theorem.

We are grateful to Karim Abadir, Paul Bekker, Hamparsum Bozdogan, Michael Browne, Frank Gerrish, Kaddour Hadri, Tõnu Kollo, Shuangzhe Liu, Daan Nel, Albert Satorra, Kazuo Shigemasu, Jos ten Berge, Peter ter Berg, Götz Trenkler, Haruo Yanai, and many others for their thoughtful and constructive comments. Of course, we welcome further comments from our readers.

Tilburg/Amsterdam — Jan R. Magnus
March 1998 — Heinz Neudecker

Preface to the third edition

Twenty years have passed since the appearance of the second edition and thirty years since the book first appeared. This is a long time, but the book still lives. Unfortunately, my coauthor Heinz Neudecker does not; he died in December 2017. Heinz was my teacher at the University of Amsterdam and I was fortunate to learn the subject of matrix calculus through differentials (then in its infancy) from his lectures and personal guidance. This technique is still a remarkably powerful tool, and Heinz Neudecker must be regarded as its founding father.

The original text of the book was written on a typewriter and then handed over to the publisher for typesetting and printing. When it came to the second edition, the typeset material could no longer be found, which is why the second edition had to be produced in an *ad hoc* manner which was not satisfactory. Many people complained about this, to me and to the publisher, and the publisher offered us to produce a new edition, freshly typeset, which would look good. In the mean time, my Russian colleagues had proposed to translate the book into Russian, and I realized that this would only be feasible if they had a good English LaTeX text. So, my secretary Josette Janssen at Tilburg University and I produced a LaTeX text with expert advice from Jozef Pijnenburg. In the process of retyping the manuscript, many small changes

were made to improve the readability and consistency of the text, but the structure of the book was not changed. The English LaTeX version was then used as the basis for the Russian edition,

Matrichnoe Differenzial'noe Ischislenie s Prilozhenijami k Statistike i Ekonometrike,

translated by my friends Anatoly Peresetsky and Pavel Katyshev, and published by Fizmatlit Publishing House, Moscow, 2002. The current third edition is based on this English LaTeX version, although I have taken the opportunity to make many improvements to the presentation of the material.

Of course, this was not the only reason for producing a third edition. It was time to take a fresh look at the material and to update the references. I felt it was appropriate to stay close to the original text, because this is the book that Heinz and I conceived and the current text is a new edition, not a new book. The main changes relative to the second edition are as follows:

- Some subjects were treated insufficiently (some of my friends would say 'incorrectly') and I have attempted to repair these omissions. This applies in particular to the discussion on matrix functions (Section 1.21), complex differentiation (Section 5.17), and Jacobians of transformations (Section 9.17).
- The text on differentiating eigenvalues and eigenvectors and associated continuity issues has been rewritten, see Sections 8.7–8.11.
- Chapter 10 has been completely rewritten, because I am now convinced that it is not useful to define Hessian matrices for vector or matrix functions. So I now define Hessian matrices only for scalar functions and for individual components of vector functions and individual elements of matrix functions. This makes life much easier.
- I have added two additional sections at the end of Chapter 17 on psychometrics, relating to correspondence analysis and linear discriminant analysis.
- Chapter 18 is new. It can be read without consulting the other chapters and provides a summary of the whole book. It can therefore be used as an introduction to matrix calculus for advanced undergraduates or Master's and PhD students in economics, statistics, mathematics, and engineering who want to know how to apply matrix calculus without going into all the theoretical details.

In addition, many small changes have been made, references have been updated, and exercises have been added. Over the past 30 years, I received many queries, problems, and requests from readers, about once every 2 weeks, which amounts to about 750 queries in 30 years. I responded to all of them and a number of these problems appear in the current text as exercises.

I am grateful to Don Andrews, Manuel Arellano, Richard Baillie, Luc Bauwens, Andrew Chesher, Gerda Claeskens, Russell Davidson, Jean-Marie

Dufour, Ronald Gallant, Eric Ghysels, Bruce Hansen, Grant Hillier, Cheng Hsiao, Guido Imbens, Guido Kuersteiner, Offer Lieberman, Esfandiar Maasoumi, Whitney Newey, Kazuhiro Ohtani, Enrique Sentana, Cezary Sielużycki, Richard Smith, Götz Trenkler, and Farshid Vahid for general encouragement and specific suggestions; to Henk Pijls for answering my questions on complex differentiation and Michel van de Velden for help on psychometric issues; to Jan Brinkhuis, Chris Muris, Franco Peracchi, Andrey Vasnev, Wendun Wang, and Yuan Yue on commenting on the new Chapter 18; to Ang Li for exceptional research assistance in updating the literature; and to Ilka van de Werve for expertly redrawing the figures. No blame attaches to any of these people in case there are remaining errors, ambiguities, or omissions; these are entirely my own responsibility, especially since I have not always followed their advice.

Cross-References. The numbering of theorems, propositions, corollaries, figures, tables, assumptions, examples, and definitions is with two digits, so that Theorem 3.5 refers to Theorem 5 in Chapter 3. Sections are numbered 1, 2,... within each chapter but always referenced with two digits so that Section 5 in Chapter 3 is referred to as Section 3.5. Equations are numbered (1), (2), ... within each chapter, and referred to with one digit if it refers to the same chapter; if it refers to another chapter we write, for example, see Equation (16) in Chapter 5. Exercises are numbered 1, 2,... after a section.

Notation. Special symbols are used to denote the derivative (matrix) D and the Hessian (matrix) H. The differential operator is denoted by d. The third edition follows the notation of earlier editions with the following exceptions. First, the symbol for the vector $(1, 1, \ldots, 1)'$ has been altered from a calligraphic s to $\imath$ (dotless i); second, the symbol i for imaginary root has been replaced by the more common i; third, $\mathrm{v}(A)$, the vector indicating the essentially distinct components of a symmetric matrix A, has been replaced by $\mathrm{vech}(A)$; fourth, the symbols for expectation, variance, and covariance (previously $\mathcal{E}$, $\mathcal{V}$, and $\mathcal{C}$) have been replaced by E, var, and cov, respectively; and fifth, we now denote the normal distribution by N (previously $\mathcal{N}$). A list of all symbols is presented in the Index of Symbols at the end of the book.

Brackets are used sparingly. We write $\operatorname{tr} A$ instead of $\operatorname{tr}(A)$, while $\operatorname{tr} AB$ denotes $\operatorname{tr}(AB)$, not $(\operatorname{tr} A)B$. Similarly, $\operatorname{vec} AB$ means $\operatorname{vec}(AB)$ and $\mathsf{d}XY$ means $\mathsf{d}(XY)$. In general, we only place brackets when there is a possibility of ambiguity.

I worked on the third edition between April and November 2018. I hope the book will continue to be useful for a few more years, and of course I welcome comments from my readers.

Amsterdam/Wapserveen Jan R. Magnus
November 2018

Note to the reader:

This preview of the the third edition of *Matrix Differential Calculus* contains only Chapter 18, the final chapter. This chapter is different from the other chapters in the book and can be read independently. It attempts to summarize matrix calculus for the user who does not want to go into all the details.

Chapter 18 is available free of charge. When quoting results from this chapter, please refer to:

Magnus, J. R. and H. Neudecker (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Third Edition, John Wiley, New York.

CHAPTER 18

Matrix calculus: the essentials

1 INTRODUCTION

This chapter differs from the other chapters in this book. It attempts to summarize the theory and the practical applications of matrix calculus in a few pages, leaving out all the subtleties that the typical user will not need. It also serves as an introduction for (advanced) undergraduates or Master's and PhD students in economics, statistics, mathematics, and engineering, who want to know how to apply matrix calculus without going into all the theoretical details. The chapter can be read independently of the rest of the book.

We begin by introducing the concept of a differential, which lies at the heart of matrix calculus. The key advantage of the differential over the more common derivative is the following. Consider the linear vector function $f(x) = Ax$ where A is an $m \times n$ matrix of constants. Then, $f(x)$ is an $m \times 1$ vector function of an $n \times 1$ vector x, and the derivative $\mathsf{D}f(x)$ is an $m \times n$ matrix (in this case, the matrix A). But the differential $\mathsf{d}f$ remains an $m \times 1$ vector. In general, the differential $\mathsf{d}f$ of a vector function $f = f(x)$ has the same dimension as f, irrespective of the dimension of the vector x, in contrast to the derivative $\mathsf{D}f(x)$.

The advantage is even larger for matrices. The differential $\mathsf{d}F$ of a matrix function $F(X)$ has the same dimension as F, irrespective of the dimension of the matrix X. The practical importance of working with differentials is huge and will be demonstrated through many examples.

We next discuss vector calculus and optimization, with and without constraints. We emphasize the importance of a correct definition and notation for the derivative, present the 'first identification theorem', which links the first differential with the first derivative, and apply these results to least squares. Then we extend the theory from vector calculus to matrix calculus and obtain the differentials of the determinant and inverse.

Matrix Differential Calculus with Applications in Statistics and Econometrics,
Third Edition. Jan R. Magnus and Heinz Neudecker.

A brief interlude on quadratic forms follows, the primary purpose of which is to show that if $x'Ax = 0$ for all x, then this does not imply that A is zero, but only that $A' = -A$. We then define the second differential and the Hessian matrix, prove the 'second identification theorem', which links the second differential with the Hessian matrix, and discuss the chain rule for second differentials. The first part of this chapter ends with four examples.

In the second (more advanced) part, we introduce the vec operator and the Kronecker product, and discuss symmetry (commutation and duplication matrices). Many examples are provided to clarify the technique. The chapter ends with an application to maximum likelihood estimation, where all elements discussed in the chapter come together.

The following notation is used. Unless specified otherwise, ϕ denotes a scalar function, f a vector function, and F a matrix function. Also, x denotes a scalar or vector argument, and X a matrix argument. All functions and variables in this chapter are real. Parentheses are used sparingly. We write $\mathsf{d}X$, $\operatorname{tr} X$, and $\operatorname{vec} X$ without parentheses, and also $\mathsf{d}XY$, $\operatorname{tr} XY$, and $\operatorname{vec} XY$ instead of $\mathsf{d}(XY)$, $\operatorname{tr}(XY)$, and $\operatorname{vec}(XY)$. However, we write $\operatorname{vech}(X)$ with parentheses for historical reasons.

2 DIFFERENTIALS

We assume that the reader is familiar with high-school calculus. This includes not only simple derivatives, such as

$$\frac{dx^2}{dx} = 2x, \qquad \frac{de^x}{dx} = e^x, \qquad \frac{d\sin x}{dx} = \cos x, \tag{1}$$

but also the chain rule, for example:

$$\frac{d(\sin x)^2}{dx} = \frac{d(\sin x)^2}{d\sin x}\,\frac{d\sin x}{dx} = 2\sin x\cos x = \sin(2x).$$

We now introduce the concept of a *differential*, by expressing (1) as

$$\mathsf{d}x^2 = 2x\,\mathsf{d}x, \qquad \mathsf{d}e^x = e^x\,\mathsf{d}x, \qquad \mathsf{d}\sin x = \cos x\,\mathsf{d}x, \tag{2}$$

where we write d rather than d to emphasize that this is a differential rather than a derivative. The two concepts are closely related, but they are not the same.

The concept of differential may be confusing for students who remember their mathematics teacher explain to them that it is wrong to view dx^2/dx as a fraction. They might wonder what $\mathsf{d}x$ and $\mathsf{d}x^2$ really are. What does $\mathsf{d}x^2 = 2x\,\mathsf{d}x$ mean? From a geometric point of view, it means that if we replace the graph of the function $\phi(x) = x^2$ at some value x by its linear approximation, that is, by the tangent line at the point (x, x^2), then an increment $\mathsf{d}x$ in x leads to an increment $\mathsf{d}x^2 = 2x\,\mathsf{d}x$ in x^2 in linear approximation. From an algebraic point of view, if we replace x by $x + \mathsf{d}x$ ('increment $\mathsf{d}x$'), then $\phi(x)$ is replaced by

$$\phi(x + \mathsf{d}x) = (x + \mathsf{d}x)^2 = x^2 + 2x\,\mathsf{d}x + (\mathsf{d}x)^2.$$

For small $\mathsf{d}x$, the term $(\mathsf{d}x)^2$ will be *very* small and, if we ignore it, we obtain the linear approximation $x^2 + 2x\,\mathsf{d}x$. The differential $\mathsf{d}x^2$ is, for a given value of x, just a function of the real variable $\mathsf{d}x$, given by the formula $\mathsf{d}x^2 = 2x\,\mathsf{d}x$.

This may sound complicated, but working with differentials is easy. The passage from (1) to (2) holds generally for any (differentiable) real-valued function ϕ, and the differential $\mathsf{d}\phi$ is thus given by the formula

$$\mathsf{d}\phi = \frac{d\phi(x)}{dx}\,\mathsf{d}x.$$

Put differently,

$$\mathsf{d}\phi = \alpha(x)\,\mathsf{d}x \iff \frac{d\phi(x)}{dx} = \alpha(x), \tag{3}$$

where α may depend on x, but not on $\mathsf{d}x$. Equation (3) is a special case of the *first identification theorem* (Theorem 18.1) in the next section. It shows that we can *identify* the derivative from the differential (and vice versa), and it shows that the new concept differential is equivalent to the familiar concept derivative. We will always work with the differential, as this has great practical advantages.

The differential is an operator, in fact a linear operator, and we have

$$\mathsf{d}a = 0, \qquad \mathsf{d}(ax) = a\,\mathsf{d}x,$$

for any scalar constant a, and

$$\mathsf{d}(x+y) = \mathsf{d}x + \mathsf{d}y, \qquad \mathsf{d}(x-y) = \mathsf{d}x - \mathsf{d}y.$$

For the product and the ratio, we have

$$\mathsf{d}(xy) = (\mathsf{d}x)y + x\,\mathsf{d}y, \qquad \mathsf{d}\left(\frac{1}{x}\right) = -\frac{\mathsf{d}x}{x^2} \qquad (x \neq 0),$$

and, in addition to the differential of the exponential function $\mathsf{d}e^x = e^x \mathsf{d}x$,

$$\mathsf{d}\log x = \frac{\mathsf{d}x}{x} \qquad (x > 0).$$

The chain rule, well-known for derivatives, also applies to differentials and is then called *Cauchy's rule of invariance*. For example,

$$\mathsf{d}(\sin x)^2 = 2\sin x\,\mathsf{d}\sin x = 2\sin x\cos x\,\mathsf{d}x = \sin(2x)\,\mathsf{d}x, \tag{4}$$

or

$$\mathsf{d}e^{x^2} = e^{x^2}\mathsf{d}x^2 = 2e^{x^2}x\,\mathsf{d}x,$$

or, combining the two previous examples,

$$\begin{aligned}\mathsf{d}e^{\sin x^2} &= e^{\sin x^2}\,\mathsf{d}\sin x^2 = e^{\sin x^2}\,\cos x^2\,\mathsf{d}x^2 \\ &= 2x\,e^{\sin x^2}\,\cos x^2\,\mathsf{d}x.\end{aligned}$$

The chain rule is a good example of the general principle that things are easier — sometimes a bit, sometimes a lot — in terms of differentials than in terms of derivatives. The chain rule in terms of differentials states that taking differentials of functions preserves composition of functions. This is easier than the chain rule in terms of derivatives. Consider, for example, the function $z = h(x) = (\sin x)^2$ as the composition of the functions

$$y = g(x) = \sin x, \qquad z = f(y) = y^2,$$

so that $h(x) = f(g(x))$. Then $\mathsf{d}y = \cos x \, \mathsf{d}x$, and hence

$$\mathsf{d}z = 2y \, \mathsf{d}y = 2y \cos x \, \mathsf{d}x = 2 \sin x \cos x \, \mathsf{d}x,$$

as expected.

The chain rule is, of course, a key instrument in differential calculus. Suppose we realize that x in (4) depends on t, say $x = t^2$. Then, we do not need to compute the differential of $(\sin t^2)^2$ all over again. We can use (4) and simply write

$$\mathsf{d}(\sin t^2)^2 = \sin(2t^2) \, \mathsf{d}t^2 = 2t \, \sin(2t^2) \, \mathsf{d}t.$$

The chain rule thus allows us to apply the rules of calculus sequentially, one after another.

In this section, we have only concerned ourselves with scalar functions of a scalar argument, and the reader may wonder why we bother to introduce differentials. They do not seem to have a great advantage over the more familiar derivatives. This is true, but when we come to vector functions of vector arguments, then the advantage will become clear.

3 VECTOR CALCULUS

Let x $(n \times 1)$ and y $(m \times 1)$ be two vectors and let y be a function of x, say $y = f(x)$. What is the derivative of y with respect to x? To help us answer this question, we first consider the linear equation

$$y = f(x) = Ax,$$

where A is an $m \times n$ matrix of constants. The derivative is A and we write

$$\frac{\partial f(x)}{\partial x'} = A. \tag{5}$$

The notation $\partial f(x)/\partial x'$ is just notation, nothing else. We sometimes write the derivative as $\mathsf{D}f(x)$ or as $\mathsf{D}f$, but we avoid the notation $f'(x)$ because this may cause confusion with the transpose. The proposed notation emphasizes that we differentiate an $m \times 1$ column vector f with respect to a $1 \times n$ row vector x', resulting in an $m \times n$ derivative matrix.

More generally, the derivative of $f(x)$ is an $m \times n$ matrix containing all partial derivatives $\partial f_i(x)/\partial x_j$, but in a specific ordering, namely

$$\frac{\partial f(x)}{\partial x'} = \begin{pmatrix} \partial f_1(x)/\partial x_1 & \partial f_1(x)/\partial x_2 & \dots & \partial f_1(x)/\partial x_n \\ \partial f_2(x)/\partial x_1 & \partial f_2(x)/\partial x_2 & \dots & \partial f_2(x)/\partial x_n \\ \vdots & \vdots & & \vdots \\ \partial f_m(x)/\partial x_1 & \partial f_m(x)/\partial x_2 & \dots & \partial f_m(x)/\partial x_n \end{pmatrix}. \tag{6}$$

There is only one definition of a vector derivative, and this is it. Of course, one can organize the mn partial derivatives in different ways, but these other combinations of the partial derivatives are not derivatives, have no practical use, and should be avoided.

Notice that each row of the derivative in (6) contains the partial derivatives of *one* element of f with respect to *all* elements of x, and that each column contains the partial derivatives of *all* elements of f with respect to *one* element of x. This is an essential characteristic of the derivative. As a consequence, the derivative of a scalar function $y = \phi(x)$, such as $y = a'x$ (where a is a vector of constants), is a row vector; in this case, a'. So the derivative of $a'x$ is a', not a.

The rules in the previous section imply that the following rules apply to vector differentials, where x and y are vectors and a is a vector of real constants, all of the same order:

$$\mathsf{d}a = 0, \qquad \mathsf{d}(x') = (\mathsf{d}x)', \qquad \mathsf{d}(a'x) = a'\mathsf{d}x,$$

$$\mathsf{d}(x+y) = \mathsf{d}x + \mathsf{d}y, \qquad \mathsf{d}(x-y) = \mathsf{d}x - \mathsf{d}y,$$

and

$$\mathsf{d}(x'y) = (\mathsf{d}x)'y + x'\mathsf{d}y.$$

Now we can see the advantage of working with differentials rather than with derivatives. When we have an $m \times 1$ vector y, which is a function of an $n \times 1$ vector of variables x, say $y = f(x)$, then the derivative is an $m \times n$ matrix, but the differential $\mathsf{d}y$ or $\mathsf{d}f$ remains an $m \times 1$ vector. This is relevant for vector functions, and even more relevant for matrix functions and for second-order derivatives, as we shall see later. The practical advantage of working with differentials is therefore that the order does not increase but always stays the same.

Corresponding to the identification result (3), we have the following relationship between the differential and the derivative.

Theorem 18.1 (first identification theorem):

$$\mathsf{d}f = A(x)\,\mathsf{d}x \iff \frac{\partial f(x)}{\partial x'} = A(x).$$

This theorem shows that there is a one-to-one correspondence between first-order differentials and first-order derivatives. In other words, the differential identifies the derivative.

Example 18.1: Consider the linear function $\phi(x) = a'x$, where a is a vector of constants. This gives

$$\mathsf{d}\phi = a'\mathsf{d}x,$$

so that the derivative is a', as we have seen before.

Example 18.2a: Next, consider the quadratic function $\phi(x) = x'Ax$, where A is a matrix of constants. Here, we have

$$\mathsf{d}\phi = (\mathsf{d}x)'Ax + x'A\,\mathsf{d}x = x'A'\mathsf{d}x + x'A\,\mathsf{d}x = x'(A + A')\,\mathsf{d}x.$$

The derivative is $x'(A + A')$, and in the special case where A is symmetric, the derivative is $2x'A$.

Now suppose that $z = f(y)$ and that $y = g(x)$, so that $z = f(g(x))$. Then,

$$\frac{\partial z}{\partial x'} = \frac{\partial z}{\partial y'}\frac{\partial y}{\partial x'}.$$

This is the chain rule for vector functions. The corresponding result for differentials is the following.

Theorem 18.2 (chain rule for first differentials): Let $z = f(y)$ and $y = g(x)$, so that $z = f(g(x))$. Then,

$$\mathsf{d}z = A(y)B(x)\,\mathsf{d}x,$$

where $A(y)$ and $B(x)$ are defined through

$$\mathsf{d}z = A(y)\,\mathsf{d}y, \qquad \mathsf{d}y = B(x)\,\mathsf{d}x.$$

Example 18.3: Let $x = (x_1, x_2, x_3)'$ and

$$f(x) = \begin{pmatrix} x_1^2 - x_2^2 \\ x_1x_2x_3 \end{pmatrix}.$$

Then, the differential is

$$\begin{aligned}\mathsf{d}f &= \begin{pmatrix} \mathsf{d}(x_1^2) - \mathsf{d}(x_2^2) \\ \mathsf{d}(x_1x_2x_3) \end{pmatrix} = \begin{pmatrix} 2x_1\mathsf{d}x_1 - 2x_2\,\mathsf{d}x_2 \\ (\mathsf{d}x_1)x_2x_3 + x_1(\mathsf{d}x_2)x_3 + x_1x_2\,\mathsf{d}x_3 \end{pmatrix} \\ &= \begin{pmatrix} 2x_1 & -2x_2 & 0 \\ x_2x_3 & x_1x_3 & x_1x_2 \end{pmatrix} \begin{pmatrix} \mathsf{d}x_1 \\ \mathsf{d}x_2 \\ \mathsf{d}x_3 \end{pmatrix},\end{aligned}$$

which identifies the derivative as

$$\frac{\partial f(x)}{\partial x'} = \begin{pmatrix} 2x_1 & -2x_2 & 0 \\ x_2x_3 & x_1x_3 & x_1x_2 \end{pmatrix}.$$

Example 18.4a: Let $x = (x_1, x_2)'$, $y = (y_1, y_2)'$, and

$$\phi(y) = e^{y_1} \sin y_2, \qquad y_1 = x_1x_2^2, \qquad y_2 = x_1^2x_2.$$

Then,

$$\mathsf{d}\phi = (\mathsf{d}e^{y_1}) \sin y_2 + e^{y_1}\, \mathsf{d} \sin y_2 = a(y)'\mathsf{d}y,$$

where

$$a(y) = e^{y_1} \begin{pmatrix} \sin y_2 \\ \cos y_2 \end{pmatrix}, \qquad \mathsf{d}y = \begin{pmatrix} \mathsf{d}y_1 \\ \mathsf{d}y_2 \end{pmatrix}.$$

Also,

$$\mathsf{d}y = \begin{pmatrix} x_2^2 & 2x_1x_2 \\ 2x_1x_2 & x_1^2 \end{pmatrix} \begin{pmatrix} \mathsf{d}x_1 \\ \mathsf{d}x_2 \end{pmatrix} = B(x)\, \mathsf{d}x.$$

Hence,

$$\mathsf{d}\phi = a(y)'\, \mathsf{d}y = a(y)'B(x)\, \mathsf{d}x = c_1\, \mathsf{d}x_1 + c_2\, \mathsf{d}x_2,$$

where

$$\begin{aligned} c_1 &= x_2e^{y_1} \left(x_2 \sin y_2 + 2x_1 \cos y_2\right), \\ c_2 &= x_1e^{y_1} \left(x_1 \cos y_2 + 2x_2 \sin y_2\right), \end{aligned}$$

so that the derivative is $\partial\phi(x)/\partial x' = (c_1, c_2)$.

4 OPTIMIZATION

Let $\phi(x)$ be a scalar differentiable function that we wish to optimize with respect to an $n \times 1$ vector x. Then we obtain the differential $\mathsf{d}\phi = a(x)'\, \mathsf{d}x$, and set $a(x) = 0$. Suppose, for example, that we wish to minimize the function

$$\phi(x) = \frac{1}{2}x'Ax - b'x, \tag{7}$$

where the matrix A is positive definite. The differential is

$$\mathsf{d}\phi = x'A\, \mathsf{d}x - b'\, \mathsf{d}x = (Ax - b)'\, \mathsf{d}x.$$

(Recall that a positive definite matrix is symmetric, by definition.) The solution $\hat{x}$ needs to satisfy $A\hat{x} - b = 0$, and hence $\hat{x} = A^{-1}b$. The function ϕ has an absolute minimum at $\hat{x}$, which can be seen by defining $y = x - \hat{x}$ and writing

$$y'Ay = (x - A^{-1}b)'A(x - A^{-1}b) = 2\phi(x) + b'A^{-1}b.$$

Since A is positive definite, $y'Ay$ has a minimum at $y = 0$ and hence $\phi(x)$ has a minimum at $x = \hat{x}$. This holds for the specific linear-quadratic function (7) and it holds more generally for any (strictly) convex function. Such functions attain a (strict) absolute minimum.

Next suppose there is a restriction, say $g(x) = 0$. Then we need to optimize subject to the restriction, and we need Lagrangian theory. This works as follows. First define the Lagrangian function, usually referred to as the Lagrangian,

$$\psi(x) = \phi(x) - \lambda g(x),$$

where λ is the Lagrange multiplier. Then we obtain the differential of ψ with respect to x,

$$\mathsf{d}\psi = \mathsf{d}\phi - \lambda \mathsf{d}g,$$

and set it equal to zero. The equations

$$\frac{\partial \phi(x)}{\partial x'} = \lambda \frac{\partial g(x)}{\partial x'}, \qquad g(x) = 0$$

are the *first-order conditions*. From these $n + 1$ equations in $n + 1$ unknowns (x and λ), we solve x and λ.

If the constraint g is a vector rather than a scalar, then we have not one but several (say, m) constraints. In that case we need m multipliers and it works like this. First, define the Lagrangian

$$\psi(x) = \phi(x) - l'g(x),$$

where $l = (\lambda_1, \lambda_2, \ldots, \lambda_m)'$ is a vector of Lagrange multipliers. Then, we obtain the differential of ψ with respect to x:

$$\mathsf{d}\psi = \mathsf{d}\phi - l'\mathsf{d}g$$

and set it equal to zero. The equations

$$\frac{\partial \phi(x)}{\partial x'} = l' \frac{\partial g(x)}{\partial x'}, \qquad g(x) = 0$$

constitute $n + m$ equations (the first-order conditions). If we can solve these equations, then we obtain the solutions, say $\hat{x}$ and $\hat{l}$.

The Lagrangian method gives necessary conditions for a local constrained extremum to occur at a given point $\hat{x}$. But how do we know that this point is in fact a maximum or a minimum? Sufficient conditions are available but they may be difficult to verify. However, in the special case where ϕ is linear-quadratic (or more generally, convex) and g is linear, ϕ attains an absolute minimum at the solution $\hat{x}$ under the constraint $g(x) = 0$.

5 LEAST SQUARES

Suppose we are given an $n \times 1$ vector y and an $n \times k$ matrix X with linearly independent columns, so that $r(X) = k$. We wish to find a $k \times 1$ vector β, such that $X\beta$ is 'as close as possible' to y in the sense that the 'error' vector $e = y - X\beta$ is minimized. A convenient scalar measure of the 'error' would be $e'e$ and our objective is to minimize

$$\phi(\beta) = \frac{e'e}{2} = \frac{(y - X\beta)'(y - X\beta)}{2}, \tag{8}$$

where we note that we write $e'e/2$ rather than $e'e$. This makes no difference, since any β which minimizes $e'e$ will also minimize $e'e/2$, but it is a common trick, useful because we know that we are minimizing a quadratic function, so that a '2' will appear in the derivative. The $1/2$ neutralizes this 2.

Differentiating ϕ in (8) gives

$$\mathsf{d}\phi = e' \,\mathsf{d}e = e' \,\mathsf{d}(y - X\beta) = -e'X \,\mathsf{d}\beta.$$

Hence, the optimum is obtained when $X'e = 0$, that is, when $X'X\hat{\beta} = X'y$, from which we obtain

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{9}$$

the least-squares solution.

If there are constraints on β, say $R\beta = r$, then we need to solve

$$\begin{array}{ll} \text{minimize} & \phi(\beta) \\ \text{subject to} & R\beta = r. \end{array}$$

We assume that the m rows of R are linearly independent, and define the Lagrangian

$$\psi(\beta) = (y - X\beta)'(y - X\beta)/2 - l'(R\beta - r),$$

where l is a vector of Lagrange multipliers. The differential is

$$\begin{aligned} \mathsf{d}\psi &= \mathsf{d}(y - X\beta)'(y - X\beta)/2 - l' \,\mathsf{d}(R\beta - r) \\ &= (y - X\beta)' \,\mathsf{d}(y - X\beta) - l'R \,\mathsf{d}\beta \\ &= -(y - X\beta)'X \,\mathsf{d}\beta - l'R \,\mathsf{d}\beta. \end{aligned}$$

Setting the differential equal to zero and denoting the restricted estimators by $\tilde{\beta}$ and $\tilde{l}$, we obtain the first-order conditions

$$(y - X\tilde{\beta})'X + \tilde{l}'R = 0, \qquad R\tilde{\beta} = r,$$

or, written differently,

$$X'X\tilde{\beta} - X'y = R'\tilde{l}, \qquad R\tilde{\beta} = r.$$

We do not know $\tilde{\beta}$ but we know $R\tilde{\beta}$. Hence, we premultiply by $R(X'X)^{-1}$. Letting $\hat{\beta} = (X'X)^{-1}X'y$ as in (9), this gives

$$r - R\hat{\beta} = R(X'X)^{-1}R'\tilde{l}.$$

Since R has full row rank, we can solve for l:

$$\tilde{l} = \left(R(X'X)^{-1}R'\right)^{-1}(r - R\hat{\beta}),$$

and hence for β:

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'\tilde{l} = \hat{\beta} + (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}(r - R\hat{\beta}).$$

Since the constraint is linear and the function ϕ is linear-quadratic as in (7), it follows that the solution $\tilde{\beta}$ indeed minimizes $\phi(\beta) = e'e/2$ under the constraint $R\beta = r$.

6 MATRIX CALCULUS

We have moved from scalar calculus to vector calculus, now we move from vector calculus to matrix calculus. When discussing matrices we assume that the reader is familiar with matrix addition and multiplication, and also knows the concepts of a determinant $|A|$ and an inverse A^{-1}. An important function of a square matrix $A = (a_{ij})$ is its *trace*, which is defined as the sum of the diagonal elements of A: $\operatorname{tr} A = \sum_i a_{ii}$. We have

$$\operatorname{tr} A = \operatorname{tr} A',$$

which is obvious because a matrix and its transpose have the same diagonal elements. Less obvious is

$$\operatorname{tr} A'B = \operatorname{tr} BA'$$

for any two matrices A and B of the same order (but not necessarily square). This follows because

$$\begin{aligned}\operatorname{tr} A'B &= \sum_j (A'B)_{jj} = \sum_j \sum_i a_{ij}b_{ij} \\ &= \sum_i \sum_j b_{ij}a_{ij} = \sum_i (BA')_{ii} = \operatorname{tr} BA'. \end{aligned} \tag{10}$$

The rules for vector differentials in Section 3 carry over to matrix differentials. Let A be a matrix of constants and let α be a scalar. Then, for any X,

$$\mathsf{d}A = 0, \qquad \mathsf{d}(\alpha X) = \alpha\,\mathsf{d}X, \qquad \mathsf{d}(X') = (\mathsf{d}X)',$$

and, for square X,

$$\mathsf{d}\operatorname{tr} X = \operatorname{tr}\mathsf{d}X.$$

If X and Y are of the same order, then

$$\mathsf{d}(X+Y) = \mathsf{d}X + \mathsf{d}Y, \qquad \mathsf{d}(X-Y) = \mathsf{d}X - \mathsf{d}Y,$$

and, if the matrix product XY is defined,

$$\mathsf{d}(XY) = (\mathsf{d}X)Y + X\mathsf{d}Y.$$

Two less trivial differentials are the determinant and the inverse. For nonsingular X we have

$$\mathsf{d}|X| = |X| \operatorname{tr} X^{-1}\, \mathsf{d}X, \tag{11}$$

and in particular, when $|X| > 0$,

$$\mathsf{d} \log |X| = \frac{\mathsf{d}|X|}{|X|} = \operatorname{tr} X^{-1}\, \mathsf{d}X.$$

The proof of (11) is a little tricky and is omitted (in this chapter, but not in Chapter 8).

The differential of the inverse is, for nonsingular X,

$$\mathsf{d}X^{-1} = -X^{-1}(\mathsf{d}X)X^{-1}. \tag{12}$$

This we can prove easily by considering the equation $X^{-1}X = I$. Differentiating both sides gives

$$(\mathsf{d}X^{-1})X + X^{-1}\mathsf{d}X = 0$$

and the result then follows by postmultiplying with X^{-1}.

The chain rule also applies to matrix functions. More precisely, if $Z = F(Y)$ and $Y = G(X)$, so that $Z = F(G(X))$, then

$$\mathsf{d}Z = A(Y)B(X)\, \mathsf{d}X,$$

where $A(Y)$ and $B(X)$ are defined through

$$\mathsf{d}Z = A(Y)\, \mathsf{d}Y, \qquad \mathsf{d}Y = B(X)\, \mathsf{d}X,$$

as in Theorem 18.2.

Regarding constrained optimization, treated for vector functions in Section 18.4, we note that this can be easily and elegantly extended to matrix constraints. If we have a matrix G (rather than a vector g) of constraints and a matrix X (rather than a vector x) of variables, then we define a matrix of multipliers $L = (\lambda_{ij})$ of the same dimension as $G = (g_{ij})$. The Lagrangian then becomes

$$\psi(X) = \phi(X) - \operatorname{tr} L'G(X),$$

where we have used the fact, also used in (10) above, that

$$\operatorname{tr} L'G = \sum_i \sum_j \lambda_{ij} g_{ij}.$$

7 INTERLUDE ON LINEAR AND QUADRATIC FORMS

Before we turn from first to second differentials, that is, from linear forms to quadratic forms, we investigate under what conditions a linear or quadratic form vanishes. The sole purpose of this section is to help the reader appreciate Theorem 18.3 in the next section.

A *linear* form is an expression such as Ax. When $Ax = 0$, this does not imply that either A or x is zero. For example, if

$$A = \begin{pmatrix} 1 & -1 \\ -2 & 2 \\ 3 & -3 \end{pmatrix}, \qquad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

then $Ax = 0$, but neither $A = 0$ nor $x = 0$.

However, when $Ax = 0$ *for every* x, then A must be zero, which can be seen by taking x to be each elementary vector e_i in turn. (The ith elementary vector is the vector with one in the ith position and zeros elsewhere.)

A *quadratic* form is an expression such as $x'Ax$. When $x'Ax = 0$, this does not imply that $A = 0$ or $x = 0$ or $Ax = 0$. Even when $x'Ax = 0$ *for every* x, it does not follow that $A = 0$, as the example

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

demonstrates. This matrix A is skew-symmetric, that is, it satisfies $A' = -A$. In fact, when $x'Ax = 0$ for every x then it follows that A *must* be skew-symmetric. This can be seen by taking $x = e_i$ which implies that $a_{ii} = 0$, and then $x = e_i + e_j$ which implies that $a_{ij} + a_{ji} = 0$.

In the special case where $x'Ax = 0$ for every x and A is symmetric, then A is both symmetric ($A' = A$) and skew-symmetric ($A' = -A$), and hence $A = 0$.

8 THE SECOND DIFFERENTIAL

The second differential is simply the differential of the first differential:

$$\mathsf{d}^2 f = \mathsf{d}(\mathsf{d}f).$$

Higher-order differentials are similarly defined, but they are seldom needed.

Example 18.2b: Let $\phi(x) = x'Ax$. Then, $\mathsf{d}\phi = x'(A + A')\,\mathsf{d}x$ and

$$\begin{aligned} \mathsf{d}^2\phi &= \mathsf{d}\left(x'(A + A')\,\mathsf{d}x\right) = (\mathsf{d}x)'(A + A')\,\mathsf{d}x + x'(A + A')\,\mathsf{d}^2 x \\ &= (\mathsf{d}x)'(A + A')\,\mathsf{d}x, \end{aligned}$$

since $\mathsf{d}^2 x = 0$.

The first differential leads to the first derivative (sometimes called the *Jacobian matrix*) and the second differential leads to the second derivative (called the *Hessian matrix*). We emphasize that the concept of Hessian matrix is only useful for scalar functions, not for vector or matrix functions. When we have a vector function f we shall consider the Hessian matrix of each element of f separately, and when we have a matrix function F we shall consider the Hessian matrix of each element of F separately.

Thus, let ϕ be a scalar function and let

$$\mathsf{d}\phi = a(x)'\mathsf{d}x, \qquad \mathsf{d}a = (\mathsf{H}\phi)\,\mathsf{d}x, \tag{13}$$

where

$$a(x)' = \frac{\partial \phi(x)}{\partial x'}, \qquad \mathsf{H}\phi = \frac{\partial a(x)}{\partial x'} = \frac{\partial}{\partial x'}\left(\frac{\partial \phi(x)}{\partial x'}\right)'.$$

The ijth element of the Hessian matrix $\mathsf{H}\phi$ is thus obtained by first calculating $a_j(x) = \partial\phi(x)/\partial x_j$ and then $(\mathsf{H}\phi)_{ij} = \partial a_j(x)/\partial x_i$. The Hessian matrix contains all second-order partial derivatives $\partial^2\phi(x)/\partial x_i\,\partial x_j$, and it is *symmetric* if ϕ is twice differentiable.

The Hessian matrix is often written as

$$\mathsf{H}\phi = \frac{\partial^2 \phi(x)}{\partial x\,\partial x'}, \tag{14}$$

where the expression on the right-hand side is a notation, the precise meaning of which is given by

$$\frac{\partial^2 \phi(x)}{\partial x\,\partial x'} = \frac{\partial}{\partial x'}\left(\frac{\partial \phi(x)}{\partial x'}\right)'. \tag{15}$$

Given (13) and using the symmetry of $\mathsf{H}\phi$, we obtain the second differential as

$$\mathsf{d}^2\phi = (\mathsf{d}a)'\,\mathsf{d}x = (\mathsf{d}x)'(\mathsf{H}\phi)\,\mathsf{d}x,$$

which shows that the second differential of ϕ is a quadratic form in $\mathsf{d}x$.

Now, suppose that we have obtained, after some calculations, that $\mathsf{d}^2\phi = (\mathsf{d}x)'B(x)\,\mathsf{d}x$. Then,

$$(\mathsf{d}x)'(\mathsf{H}\phi - B(x))\,\mathsf{d}x = 0$$

for all $\mathsf{d}x$. Does this imply that $\mathsf{H}\phi = B(x)$? No, it does not, as we have seen in the previous section. It does, however, imply that

$$(\mathsf{H}\phi - B(x))' + (\mathsf{H}\phi - B(x)) = 0,$$

and hence that $\mathsf{H}\phi = (B(x) + B(x)')/2$, using the symmetry of $\mathsf{H}\phi$. This proves the following result.

Theorem 18.3 (second identification theorem):

$$\mathsf{d}^2\phi = (\mathsf{d}x)'B(x)\,\mathsf{d}x \iff \mathsf{H}\phi = \frac{B(x) + B(x)'}{2}.$$

The second identification theorem shows that there is a one-to-one correspondence between second-order differentials and second-order derivatives, but only if we make the matrix $B(x)$ in the quadratic form symmetric. Hence, the second differential identifies the second derivative.

Example 18.2c: Consider again the quadratic function $\phi(x) = x'Ax$. Then we can start with $\mathsf{d}\phi = x'(A + A')\,\mathsf{d}x$, as in Example 18.2b, and obtain $\mathsf{d}^2\phi = (\mathsf{d}x)'(A + A')\,\mathsf{d}x$. The matrix in the quadratic form is already symmetric, so we obtain directly $\mathsf{H}\phi = A + A'$.

Alternatively — and this is often quicker — we differentiate ϕ twice without writing out the first differential in its final form. From

$$\mathsf{d}\phi = (\mathsf{d}x)'Ax + x'A\,\mathsf{d}x,$$

we thus obtain

$$\mathsf{d}^2\phi = 2(\mathsf{d}x)'A\,\mathsf{d}x, \tag{16}$$

which identifies the Hessian matrix as $\mathsf{H}\phi = A + A'$. (Notice that the matrix A in (16) is not necessarily symmetric.)

Even with such a simple function as $\phi(x) = x'Ax$, the advantage and elegance of using differentials is clear. Without differentials we would need to prove first that $\partial a'x/\partial x' = a'$ and $\partial x'Ax/\partial x' = x'(A + A')$, and then use (15) to obtain

$$\frac{\partial^2 x'Ax}{\partial x\,\partial x'} = \frac{\partial(x'(A + A'))'}{\partial x'} = \frac{\partial(A + A')x}{\partial x'} = A + A',$$

which is cumbersome in this simple case and not practical in more complex situations.

9 CHAIN RULE FOR SECOND DIFFERENTIALS

Let us now return to Example 18.2b. The function ϕ in this example is a function of x, and x is the argument of interest. This is why $\mathsf{d}^2x = 0$. But if ϕ is a function of x, which in turn is a function of t, then it is no longer true that d^2x equals zero. More generally, suppose that $z = f(y)$ and that $y = g(x)$, so that $z = f(g(x))$. Then,

$$\mathsf{d}z = A(y)\,\mathsf{d}y$$

and

$$\mathsf{d}^2z = (\mathsf{d}A)\,\mathsf{d}y + A(y)\,\mathsf{d}^2y. \tag{17}$$

This is true whether or not y depends on some other variables. If we think of z as a function of y, then $\mathsf{d}^2y = 0$, but if y depends on x then d^2y is not zero; in fact,

$$\mathsf{d}y = B(x)\,\mathsf{d}x, \qquad \mathsf{d}^2y = (\mathsf{d}B)\,\mathsf{d}x.$$

This gives us the following result.

Theorem 18.4 (chain rule for second differentials): Let $z = f(y)$ and $y = g(x)$, so that $z = f(g(x))$. Then,

$$\mathsf{d}^2 z = (\mathsf{d}A)B(x)\,\mathsf{d}x + A(y)(\mathsf{d}B)\,\mathsf{d}x,$$

where $A(y)$ and $B(x)$ are defined through

$$\mathsf{d}z = A(y)\,\mathsf{d}y, \qquad \mathsf{d}y = B(x)\,\mathsf{d}x.$$

In practice, one usually avoids Theorem 18.4 by going back to the first differential $\mathsf{d}z = A(y)\,\mathsf{d}y$ and differentiating again. This gives (17), from which we obtain the result step by step.

Example 18.4b: Let

$$\phi(y_1, y_2) = e^{y_1}\sin y_2, \qquad y_1 = x_1 x_2^2, \qquad y_2 = x_1^2 x_2.$$

Then, by Theorem 18.4,

$$\mathsf{d}^2\phi = (\mathsf{d}a)'B(x)\,\mathsf{d}x + a(y)'(\mathsf{d}B)\,\mathsf{d}x,$$

where

$$a(y) = e^{y_1}\begin{pmatrix}\sin y_2\\ \cos y_2\end{pmatrix}, \qquad B(x) = \begin{pmatrix}x_2^2 & 2x_1x_2\\ 2x_1x_2 & x_1^2\end{pmatrix}.$$

Now, letting

$$C(y) = e^{y_1}\begin{pmatrix}\sin y_2 & \cos y_2\\ \cos y_2 & -\sin y_2\end{pmatrix}$$

and

$$D_1(x) = 2\begin{pmatrix}0 & x_2\\ x_2 & x_1\end{pmatrix}, \qquad D_2(x) = 2\begin{pmatrix}x_2 & x_1\\ x_1 & 0\end{pmatrix},$$

we obtain

$$\mathsf{d}a = C(y)\,\mathsf{d}y = C(y)B(x)\,\mathsf{d}x$$

and

$$\mathsf{d}B = (\mathsf{d}x_1)D_1(x) + (\mathsf{d}x_2)D_2(x).$$

It is convenient to write $\mathsf{d}x_1$ and $\mathsf{d}x_2$ in terms of $\mathsf{d}x$, which can be done by defining $e_1 = (1,0)'$ and $e_2 = (0,1)'$. Then, $\mathsf{d}x_1 = e_1'\mathsf{d}x$ and $\mathsf{d}x_2 = e_2'\mathsf{d}x$, and hence

$$\begin{aligned}\mathsf{d}^2\phi &= (\mathsf{d}a)'B(x)\,\mathsf{d}x + a(y)'(\mathsf{d}B)\,\mathsf{d}x\\ &= (\mathsf{d}x)'B(x)C(y)B(x)\,\mathsf{d}x + a(y)'\left((\mathsf{d}x_1)D_1(x) + (\mathsf{d}x_2)D_2(x)\right)\,\mathsf{d}x\\ &= (\mathsf{d}x)'B(x)C(y)B(x)\,\mathsf{d}x + (\mathsf{d}x)'e_1a(y)'D_1(x)\,\mathsf{d}x + (\mathsf{d}x)'e_2a(y)'D_2(x)\,\mathsf{d}x\\ &= (\mathsf{d}x)'\left[B(x)C(y)B(x) + e_1a(y)'D_1(x) + e_2a(y)'D_2(x)\right]\,\mathsf{d}x.\end{aligned}$$

Some care is required where to position the scalars $e_1'\mathsf{d}x$ and $e_2'\mathsf{d}x$ in the matrix product. A scalar can be positioned anywhere in a matrix product, but we wish to position the two scalars in such a way that the usual matrix multiplication rules still apply.

Having obtained the second differential in the desired form, Theorem 18.3 implies that the Hessian is equal to

$$\begin{aligned}\mathsf{H}\phi = B(x)C(y)B(x) &+ \frac{1}{2}\left(e_1 a(y)' D_1(x) + D_1(x)a(y)e_1'\right) \\ &+ \frac{1}{2}\left(e_2 a(y)' D_2(x) + D_2(x)a(y)e_2'\right).\end{aligned}$$

10 FOUR EXAMPLES

Let us provide four examples to show how the second differential can be obtained. The first three examples relate to scalar functions and the fourth example to a matrix function. The matrix X has order $n \times q$ in Examples 18.5a and 18.6a, and order $n \times n$ in Examples 18.7a and 18.8a.

Example 18.5a: Let $\phi(X) = \operatorname{tr} X'AX$. Then,

$$\begin{aligned}\mathsf{d}\phi &= \mathsf{d}(\operatorname{tr} X'AX) = \operatorname{tr}\mathsf{d}(X'AX) \\ &= \operatorname{tr}(\mathsf{d}X)'AX + \operatorname{tr} X'A\,\mathsf{d}X = \operatorname{tr} X'(A + A')\,\mathsf{d}X\end{aligned}$$

and

$$\mathsf{d}^2\phi = \mathsf{d}\operatorname{tr} X'(A + A')\,\mathsf{d}X = \operatorname{tr}(\mathsf{d}X)'(A + A')\,\mathsf{d}X.$$

Example 18.6a: Let $\phi(X) = \log|X'X|$. Then,

$$\begin{aligned}\mathsf{d}\phi &= \mathsf{d}\log|X'X| = \operatorname{tr}(X'X)^{-1}\mathsf{d}(X'X) \\ &= \operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'X + \operatorname{tr}(X'X)^{-1}X'\mathsf{d}X = 2\operatorname{tr}(X'X)^{-1}X'\mathsf{d}X\end{aligned}$$

and

$$\begin{aligned}\mathsf{d}^2\phi &= 2\,\mathsf{d}\left(\operatorname{tr}(X'X)^{-1}X'\mathsf{d}X\right) \\ &= 2\operatorname{tr}(\mathsf{d}(X'X)^{-1})X'\mathsf{d}X + 2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'\mathsf{d}X \\ &= -2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X'X)(X'X)^{-1}X'\mathsf{d}X + 2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'\mathsf{d}X \\ &= -2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'X(X'X)^{-1}X'\mathsf{d}X \\ &\quad - 2\operatorname{tr}(X'X)^{-1}X'(\mathsf{d}X)(X'X)^{-1}X'\mathsf{d}X + 2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'\mathsf{d}X \\ &= 2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'M\mathsf{d}X - 2\operatorname{tr}(X'X)^{-1}X'(\mathsf{d}X)(X'X)^{-1}X'\mathsf{d}X,\end{aligned}$$

where $M = I_n - X(X'X)^{-1}X'$. Let us explain some of the steps in more detail. The second equality follows from considering $(X'X)^{-1}X'\mathsf{d}X$ as a product

of three matrices: $(X'X)^{-1}$, X', and $\mathsf{d}X$ (a matrix of constants), the third equality uses the differential of the inverse in (12), and the fourth equality separates $\mathsf{d}X'X$ into $(\mathsf{d}X)'X + X'\mathsf{d}X$.

Example 18.7a: Let $\phi(X) = \operatorname{tr} X^k$ for $k = 1, 2, \ldots$ Then, for $k \geq 1$,

$$\begin{aligned}\mathsf{d}\phi &= \operatorname{tr}(\mathsf{d}X)X^{k-1} + \operatorname{tr} X(\mathsf{d}X)X^{k-2} + \cdots + \operatorname{tr} X^{k-1}\mathsf{d}X \\ &= k \operatorname{tr} X^{k-1}\mathsf{d}X,\end{aligned}$$

and for $k \geq 2$,

$$\mathsf{d}^2\phi = k \operatorname{tr}(\mathsf{d}X^{k-1})\,\mathsf{d}X = k \sum_{j=0}^{k-2} \operatorname{tr} X^j(\mathsf{d}X)X^{k-2-j}\mathsf{d}X.$$

Example 18.8a: Let $F(X) = AX^{-1}B$. Then,

$$\mathsf{d}F = A(\mathsf{d}X^{-1})B = -AX^{-1}(\mathsf{d}X)X^{-1}B$$

and

$$\begin{aligned}\mathsf{d}^2F &= -A(\mathsf{d}X^{-1})(\mathsf{d}X)X^{-1}B - AX^{-1}(\mathsf{d}X)(\mathsf{d}X^{-1})B \\ &= 2AX^{-1}(\mathsf{d}X)X^{-1}(\mathsf{d}X)X^{-1}B.\end{aligned}$$

These four examples provide the second differential; they do not yet provide the Hessian matrix. In Section 18.14, we shall discuss the same four examples and obtain the Hessian matrices.

11 THE KRONECKER PRODUCT AND VEC OPERATOR

The theory and the four examples in the previous two sections demonstrate the elegance and simplicity of obtaining first and second differentials of scalar, vector, and matrix functions. But we also want to relate these first and second differentials to Jacobian matrices (first derivatives) and Hessian matrices (second derivatives). For this we need some more machinery, namely the vec operator and the Kronecker product.

First, the vec operator. Consider an $m \times n$ matrix A. This matrix has n columns, say $a_1, \ldots, a_n$. Now define the $mn \times 1$ vector $\operatorname{vec} A$ as the vector which stacks these columns one underneath the other:

$$\operatorname{vec} A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}.$$

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

then $\operatorname{vec} A = (1, 4, 2, 5, 3, 6)'$. Of course, we have

$$\mathsf{d} \operatorname{vec} X = \operatorname{vec} \mathsf{d}X. \tag{18}$$

If A and B are matrices of the same order, then we know from (10) that $\operatorname{tr} A'B = \sum_{ij} a_{ij} b_{ij}$. But $(\operatorname{vec} A)'(\operatorname{vec} B)$ is also equal to this double sum. Hence,

$$\operatorname{tr} A'B = (\operatorname{vec} A)'(\operatorname{vec} B), \tag{19}$$

an important equality linking the vec operator to the trace.

We also need the Kronecker product. Let A be an $m \times n$ matrix and B a $p \times q$ matrix. The $mp \times nq$ matrix defined by

$$\begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}$$

is called the *Kronecker product* of A and B and is written as $A \otimes B$. The Kronecker product $A \otimes B$ is thus defined for any pair of matrices A and B, unlike the matrix product AB which exists only if the number of columns in A equals the number of rows in B or if either A or B is a scalar.

The following three properties justify the name Kronecker *product*:

$$A \otimes B \otimes C = (A \otimes B) \otimes C = A \otimes (B \otimes C),$$
$$(A + B) \otimes (C + D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D,$$

if A and B have the same order and C and D have the same order (not necessarily equal to the order of A and B), and

$$(A \otimes B)(C \otimes D) = AC \otimes BD,$$

if AC and BD exist.

The transpose of a Kronecker product is

$$(A \otimes B)' = A' \otimes B'.$$

If A and B are square matrices (not necessarily of the same order), then

$$\operatorname{tr}(A \otimes B) = (\operatorname{tr} A)(\operatorname{tr} B),$$

and if A and B are nonsingular, then

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

The Kronecker product and the vec operator are related through the equality

$$\operatorname{vec} ab' = b \otimes a,$$

where a and b are column vectors of arbitrary order. Using this inequality, we see that

$$\begin{aligned}\operatorname{vec}(Abe'C) &= \operatorname{vec}(Ab)(C'e)' = (C'e) \otimes (Ab) \\ &= (C' \otimes A)(e \otimes b) = (C' \otimes A) \operatorname{vec} be'\end{aligned}$$

for any vectors b and e. Then, writing $B = \sum_j b_j e_j'$ where b_j and e_j denote the jth column of B and I, respectively, we obtain the following important relationship, which is used frequently.

Theorem 18.5: For any matrices A, B, and C for which the product ABC is defined, we have

$$\operatorname{vec} ABC = (C' \otimes A) \operatorname{vec} B.$$

12 IDENTIFICATION

When we move from vector calculus to matrix calculus, we need an ordering of the functions and of the variables. It does not matter *how* we order them (any ordering will do), but an ordering is essential. We want to define matrix derivatives within the established theory of vector derivatives in such a way that trivial changes such as relabeling functions or variables have only trivial consequences for the derivative: rows and columns are permuted, but the rank is unchanged and the determinant (in the case of a square matrix) is also unchanged, apart possibly from its sign. This is what we need to achieve. The arrangement of the partial derivatives matters, because a derivative is more than just a collection of partial derivatives. It is a mathematical concept, a mathematical unit.

Thus motivated, we shall view the matrix function $F(X)$ as a vector function $f(x)$, where $f = \operatorname{vec} F$ and $x = \operatorname{vec} X$. We then obtain the following extension of the first identification theorem:

$$\mathsf{d} \operatorname{vec} F = A(X)\, \mathsf{d} \operatorname{vec} X \iff \frac{\partial \operatorname{vec} F(X)}{\partial (\operatorname{vec} X)'} = A(X),$$

and, similarly, for the second identification theorem:

$$\mathsf{d}^2 \phi = (\mathsf{d} \operatorname{vec} X)' B(X)\, \mathsf{d} \operatorname{vec} X \iff \mathsf{H}\phi = \frac{B(X) + B(X)'}{2},$$

where we notice, as in Section 18.8, that we only provide the Hessian matrix for scalar functions, not for vector or matrix functions.

13 THE COMMUTATION MATRIX

At this point, we need to introduce the commutation matrix. Let A be an $m \times n$ matrix. The vectors $\operatorname{vec} A$ and $\operatorname{vec} A'$ contain the same mn elements, but in a different order. Hence, there exists a unique $mn \times mn$ matrix, which transforms $\operatorname{vec} A$ into $\operatorname{vec} A'$. This matrix contains mn ones and $mn(mn-1)$ zeros and is called the *commutation matrix*, denoted by K_{mn}. (If $m = n$, we write K_n instead of K_{nn}.) Thus,

$$K_{mn} \operatorname{vec} A = \operatorname{vec} A'. \tag{20}$$

It can be shown that K_{mn} is orthogonal, i.e. $K'_{mn} = K_{mn}^{-1}$. Also, premultiplying (20) by K_{nm} gives $K_{nm}K_{mn} \operatorname{vec} A = \operatorname{vec} A$, which shows that $K_{nm}K_{mn} = I_{mn}$. Hence,

$$K'_{mn} = K_{mn}^{-1} = K_{nm}.$$

The key property of the commutation matrix enables us to interchange (commute) the two matrices of a Kronecker product:

$$K_{pm}(A \otimes B) = (B \otimes A)K_{qn} \tag{21}$$

for any $m \times n$ matrix A and $p \times q$ matrix B. This is easiest shown, not by proving a matrix identity but by proving that the *effect* of the two matrices on an arbitrary vector is the same. Thus, let X be an arbitrary $q \times n$ matrix. Then, by repeated application of (20) and Theorem 18.5,

$$\begin{aligned} K_{pm}(A \otimes B) \operatorname{vec} X &= K_{pm} \operatorname{vec} BXA' = \operatorname{vec} AX'B' \\ &= (B \otimes A) \operatorname{vec} X' = (B \otimes A)K_{qn} \operatorname{vec} X. \end{aligned}$$

Since X is arbitrary, (21) follows.

The commutation matrix has many applications in matrix theory. Its importance in matrix calculus stems from the fact that it transforms $\mathsf{d} \operatorname{vec} X'$ into $\mathsf{d} \operatorname{vec} X$. The simplest example is the matrix function $F(X) = X'$, where X is an $n \times q$ matrix. Then,

$$\mathsf{d} \operatorname{vec} F = \operatorname{vec} \mathsf{d}X' = K_{nq} \operatorname{vec} \mathsf{d}X,$$

so that the derivative is $\mathsf{D} \operatorname{vec} F = K_{nq}$.

The commutation matrix is also essential in identifying the Hessian matrix from the second differential. The second differential of a scalar function often takes the form of a trace, either $\operatorname{tr} A(\mathsf{d}X)'B\mathsf{d}X$ or $\operatorname{tr} A(\mathsf{d}X)B\mathsf{d}X$. We then have the following result, based on (19) and Theorem 18.5.

Theorem 18.6: Let ϕ be a twice differentiable real-valued function of an $n \times q$ matrix X. Then,

$$\mathsf{d}^2\phi = \operatorname{tr} A(\mathsf{d}X)'B\mathsf{d}X \iff \mathsf{H}\phi = \frac{1}{2}(A' \otimes B + A \otimes B')$$

and

$$\mathsf{d}^2\phi = \operatorname{tr} A(\mathsf{d}X)B\mathsf{d}X \iff \mathsf{H}\phi = \frac{1}{2}K_{qn}(A' \otimes B + B' \otimes A).$$

To identify the Hessian matrix from the first expression, we do not need the commutation matrix, but we do need the commutation matrix to identify the Hessian matrix from the second expression.

14 FROM SECOND DIFFERENTIAL TO HESSIAN

We continue with the same four examples as discussed in Section 18.10, showing how to obtain the Hessian matrices from the second differentials, using Theorem 18.6.

Example 18.5b: Let $\phi(X) = \operatorname{tr} X'AX$, where X is an $n \times q$ matrix. Then,

$$\mathsf{d}\phi = \operatorname{tr} X'(A + A')\,\mathsf{d}X = \operatorname{tr} C'\mathsf{d}X = (\operatorname{vec} C)'\mathsf{d}\operatorname{vec} X,$$

where $C = (A + A')X$, and $\mathsf{d}^2\phi = \operatorname{tr}(\mathsf{d}X)'(A + A')\,\mathsf{d}X$. Hence, the derivative is $\mathsf{D}\phi = (\operatorname{vec} C)'$ and the Hessian is $\mathsf{H}\phi = I_q \otimes (A + A')$.

Example 18.6b: Let $\phi(X) = \log|X'X|$, where X is an $n \times q$ matrix of full column rank. Then, letting $C = X(X'X)^{-1}$ and $M = I_n - X(X'X)^{-1}X'$,

$$\mathsf{d}\phi = 2\operatorname{tr} C'\mathsf{d}X = 2(\operatorname{vec} C)'\mathsf{d}\operatorname{vec} X$$

and

$$\mathsf{d}^2\phi = 2\operatorname{tr}(X'X)^{-1}(\mathsf{d}X)'M\mathsf{d}X - 2\operatorname{tr} C'(\mathsf{d}X)C'\mathsf{d}X.$$

This gives $\mathsf{D}\phi = 2(\operatorname{vec} C)'$ and

$$\mathsf{H}\phi = 2(X'X)^{-1} \otimes M - 2K_{qn}(C \otimes C').$$

Example 18.7b: Let $\phi(X) = \operatorname{tr} X^k$ for $k = 1, 2, \ldots$, where X is a square $n \times n$ matrix. Then, for $k \geq 1$,

$$\mathsf{d}\phi = k\operatorname{tr} X^{k-1}\mathsf{d}X = k(\operatorname{vec} X'^{k-1})'\mathsf{d}\operatorname{vec} X,$$

and for $k \geq 2$,

$$\mathsf{d}^2\phi = k\sum_{j=0}^{k-2} \operatorname{tr} X^j(\mathsf{d}X)X^{k-2-j}\mathsf{d}X.$$

This gives $\mathsf{D}\phi = k(\operatorname{vec} X'^{k-1})'$ and

$$\mathsf{H}\phi = (k/2)\sum_{j=0}^{k-2} K_n(X'^j \otimes X^{k-2-j} + X'^{k-2-j} \otimes X^j).$$

Example 18.8b: Let $F(X) = AX^{-1}B$, where X is a nonsingular $n \times n$ matrix. Then, using Theorem 18.5,

$$\mathsf{d}\,\mathrm{vec}\,F = -\left((X^{-1}B)' \otimes (AX^{-1})\right)\mathsf{d}\,\mathrm{vec}\,X,$$

and hence

$$\mathsf{D}\,\mathrm{vec}\,F = -(X^{-1}B)' \otimes (AX^{-1}).$$

To obtain the Hessian matrix of the stth element of F, we let

$$C_{ts} = X^{-1}Be_te_s'AX^{-1},$$

where e_s and e_t are elementary vectors with 1 in the sth (respectively, tth) position and zeros elsewhere. Then,

$$\mathsf{d}^2F_{st} = 2e_s'AX^{-1}(\mathsf{d}X)X^{-1}(\mathsf{d}X)X^{-1}Be_t = 2\,\mathrm{tr}\,C_{ts}(\mathsf{d}X)X^{-1}(\mathsf{d}X)$$

and hence

$$\mathsf{H}F_{st} = K_n(C_{ts}' \otimes X^{-1} + X'^{-1} \otimes C_{ts}).$$

15 SYMMETRY AND THE DUPLICATION MATRIX

Many matrices in statistics and econometrics are symmetric, for example variance matrices. When we differentiate with respect to symmetric matrices, we must take the symmetry into account and we need the duplication matrix.

Let A be a square $n \times n$ matrix. Then $\mathrm{vech}(A)$ will denote the $\frac{1}{2}n(n+1) \times 1$ vector that is obtained from $\mathrm{vec}\,A$ by eliminating all elements of A above the diagonal. For example, for $n = 3$,

$$\mathrm{vec}\,A = (a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32}, a_{13}, a_{23}, a_{33})'$$

and

$$\mathrm{vech}(A) = (a_{11}, a_{21}, a_{31}, a_{22}, a_{32}, a_{33})'. \tag{22}$$

In this way, for symmetric A, $\mathrm{vech}(A)$ contains only the generically distinct elements of A. Since the elements of $\mathrm{vec}\,A$ are those of $\mathrm{vech}(A)$ with some repetitions, there exists a unique $n^2 \times \frac{1}{2}n(n+1)$ matrix which transforms, for symmetric A, $\mathrm{vech}(A)$ into $\mathrm{vec}\,A$. This matrix is called the *duplication matrix* and is denoted by D_n. Thus,

$$D_n\,\mathrm{vech}(A) = \mathrm{vec}\,A \qquad (A = A'). \tag{23}$$

The matrix D_n has full column rank $\frac{1}{2}n(n+1)$, so that $D_n'D_n$ is nonsingular. This implies that $\mathrm{vech}(A)$ can be uniquely solved from (23), and we have

$$\mathrm{vech}(A) = (D_n'D_n)^{-1}D_n'\,\mathrm{vec}\,A \qquad (A = A').$$

One can show (but we will not do so here) that the duplication matrix is connected to the commutation matrix by

$$K_n D_n = D_n, \qquad D_n(D_n'D_n)^{-1}D_n' = \frac{1}{2}(I_{n^2} + K_n).$$

Much of the interest in the duplication matrix is due to the importance of the matrix $D_n'(A \otimes A)D_n$, where A is an $n \times n$ matrix. This matrix is important, because the scalar function $\phi(X) = \operatorname{tr} AX'AX$ occurs frequently in statistics and econometrics, for example in the next section on maximum likelihood. When A and X are known to be symmetric we have

$$\begin{aligned}\mathsf{d}^2\phi &= 2\operatorname{tr} A(\mathsf{d}X)'A\,\mathsf{d}X = 2(\mathsf{d}\operatorname{vec} X)'(A \otimes A)\mathsf{d}\operatorname{vec} X \\ &= 2(\mathsf{d}\operatorname{vech}(X))'D_n'(A \otimes A)D_n\,\mathsf{d}\operatorname{vech}(X),\end{aligned}$$

and hence, $\mathsf{H}\phi = 2D_n'(A \otimes A)D_n$.

From the relationship (again not proved here)

$$D_n(D_n'D_n)^{-1}D_n'(A \otimes A)D_n = (A \otimes A)D_n,$$

which is valid for any $n \times n$ matrix A, not necessarily symmetric, we obtain the inverse

$$(D_n'(A \otimes A)D_n)^{-1} = (D_n'D_n)^{-1}D_n'(A^{-1} \otimes A^{-1})D_n(D_n'D_n)^{-1}, \tag{24}$$

where A is nonsingular. Finally, we present the determinant:

$$|D_n'(A \otimes A)D_n| = 2^{\frac{1}{2}n(n-1)}|A|^{n+1}. \tag{25}$$

16 MAXIMUM LIKELIHOOD

This final section brings together most of the material that has been treated in this chapter: first and second differentials, the Hessian matrix, and the treatment of symmetry (duplication matrix).

We consider a sample of $m \times 1$ vectors $y_1, y_2, \ldots, y_n$ from the multivariate normal distribution with mean μ and variance Ω, where Ω is positive definite and $n \geq m + 1$. The density of y_i is

$$f(y_i) = (2\pi)^{-m/2}|\Omega|^{-1/2}\exp\left(-\frac{1}{2}(y_i - \mu)'\Omega^{-1}(y_i - \mu)\right),$$

and since the y_i are independent and identically distributed, the joint density of $(y_1, \ldots, y_n)$ is given by $\prod_i f(y_i)$. The 'likelihood' is equal to the joint density, but now thought of as a function of the parameters μ and Ω, rather than of the observations. Its logarithm is the 'loglikelihood', which here takes the form

$$\Lambda(\mu, \Omega) = -\frac{mn}{2}\log 2\pi - \frac{n}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)'\Omega^{-1}(y_i - \mu).$$

The maximum likelihood estimators are obtained by maximizing the loglikelihood (which is the same, but usually easier, as maximizing the likelihood). Thus, we differentiate Λ and obtain

$$\begin{aligned}\mathsf{d}\Lambda &= -\frac{n}{2}\,\mathsf{d}\log|\Omega| + \frac{1}{2}\sum_{i=1}^{n}(\mathsf{d}\mu)'\Omega^{-1}(y_i-\mu)\\ &\quad -\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'(\mathsf{d}\Omega^{-1})(y_i-\mu) + \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Omega^{-1}\mathsf{d}\mu\\ &= -\frac{n}{2}\,\mathsf{d}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'(\mathsf{d}\Omega^{-1})(y_i-\mu) + \sum_{i=1}^{n}(y_i-\mu)'\Omega^{-1}\mathsf{d}\mu\\ &= -\frac{n}{2}\operatorname{tr}(\Omega^{-1}\mathsf{d}\Omega + S\mathsf{d}\Omega^{-1}) + \sum_{i=1}^{n}(y_i-\mu)'\Omega^{-1}\mathsf{d}\mu, \end{aligned} \tag{26}$$

where

$$S = S(\mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)'.$$

Denoting the maximum likelihood estimators by $\hat{\mu}$ and $\hat{\Omega}$, letting $\hat{S} = S(\hat{\mu})$, and setting $\mathsf{d}\Lambda = 0$ then implies that

$$\operatorname{tr}\left(\hat{\Omega}^{-1} - \hat{\Omega}^{-1}\hat{S}\hat{\Omega}^{-1}\right)\mathsf{d}\Omega = 0$$

for all $\mathsf{d}\Omega$ and

$$\sum_{i=1}^{n}(y_i-\hat{\mu})'\hat{\Omega}^{-1}\mathsf{d}\mu = 0$$

for all $\mathsf{d}\mu$. This, in turn, implies that

$$\hat{\Omega}^{-1} = \hat{\Omega}^{-1}\hat{S}\hat{\Omega}^{-1}, \qquad \sum_{i=1}^{n}(y_i-\hat{\mu}) = 0.$$

Hence, the maximum likelihood estimators are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{27}$$

and

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})'. \tag{28}$$

We note that the condition that Ω is symmetric has not been imposed. But since the solution (28) is symmetric, imposing the condition would have made no difference.

The second differential is obtained by differentiating (26) again. This gives

$$\begin{aligned}\mathsf{d}^2\Lambda = &-\frac{n}{2}\operatorname{tr}\left((\mathsf{d}\Omega^{-1})\mathsf{d}\Omega + (\mathsf{d}S)\mathsf{d}\Omega^{-1} + S\mathsf{d}^2\Omega^{-1}\right) - n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu \\ &+ \sum_{i=1}^{n}(y_i-\mu)'(\mathsf{d}\Omega^{-1})\mathsf{d}\mu. \end{aligned} \tag{29}$$

We are usually not primarily interested in the Hessian matrix but in its expectation. Hence, we do not evaluate (29) further and first take expectations. Since $\mathrm{E}(S) = \Omega$ and $\mathrm{E}(\mathsf{d}S) = 0$, we obtain

$$\begin{aligned}\mathrm{E}\,\mathsf{d}^2\Lambda &= -\frac{n}{2}\operatorname{tr}\left((\mathsf{d}\Omega^{-1})\mathsf{d}\Omega + \Omega\mathsf{d}^2\Omega^{-1}\right) - n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu \\ &= \frac{n}{2}\operatorname{tr}\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1}\mathsf{d}\Omega - n\operatorname{tr}(\mathsf{d}\Omega)\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1} - n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu \\ &= -\frac{n}{2}\operatorname{tr}\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1}\mathsf{d}\Omega - n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu, \end{aligned} \tag{30}$$

using the facts that $\mathsf{d}\Omega^{-1} = -\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1}$ and

$$\begin{aligned}\mathsf{d}^2\Omega^{-1} &= -(\mathsf{d}\Omega^{-1})(\mathsf{d}\Omega)\Omega^{-1} - \Omega^{-1}(\mathsf{d}\Omega)\mathsf{d}\Omega^{-1} \\ &= 2\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1}.\end{aligned}$$

To obtain the 'information matrix' we need to take the symmetry of Ω into account and this is where the duplication matrix appears. So far, we have avoided the vec operator and in practical situations one should work with differentials (rather than with derivatives) as long as possible. But we cannot go further than (30) without use of the vec operator. Thus, from (30),

$$\begin{aligned}-\mathrm{E}\mathsf{d}^2\Lambda &= \frac{n}{2}\operatorname{tr}\Omega^{-1}(\mathsf{d}\Omega)\Omega^{-1}\mathsf{d}\Omega + n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu \\ &= \frac{n}{2}(\mathsf{d}\operatorname{vec}\Omega)'(\Omega^{-1}\otimes\Omega^{-1})\,\mathsf{d}\operatorname{vec}\Omega + n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu \\ &= \frac{n}{2}(\mathsf{d}\operatorname{vech}(\Omega))'D_m'(\Omega^{-1}\otimes\Omega^{-1})D_m\,\mathsf{d}\operatorname{vech}(\Omega) + n(\mathsf{d}\mu)'\Omega^{-1}\mathsf{d}\mu.\end{aligned}$$

Hence, the information matrix for μ and $\operatorname{vech}(\Omega)$ is

$$\mathcal{F} = n\begin{pmatrix}\Omega^{-1} & 0\\ 0 & \frac{1}{2}D_m'(\Omega^{-1}\otimes\Omega^{-1})D_m\end{pmatrix}.$$

The results on the duplication matrix in Section 18.15 also allow us to obtain the inverse:

$$(\mathcal{F}/n)^{-1} = \begin{pmatrix}\Omega & 0\\ 0 & 2(D_m'D_m)^{-1}D_m'(\Omega\otimes\Omega)D_m(D_m'D_m)^{-1}\end{pmatrix}$$

and the determinant:

$$|\mathcal{F}/n| = |\Omega|\cdot|2(D_m'D_m)^{-1}D_m'(\Omega\otimes\Omega)D_m(D_m'D_m)^{-1}| = 2^m|\Omega|^{m+2}.$$

FURTHER READING

§2–3. Chapter 5 discusses differentials in more detail, and contains the first identification theorem (Theorem 5.6) and the chain rule for first differentials (Theorem 5.9), officially called 'Cauchy's rule of invariance'.

§4. Optimization is discussed in Chapter 7.

§5. See Chapter 11, Sections 11.29–11.32 and Chapter 13, Sections 13.4 and 13.19.

§6. The trace is discussed in Section 1.10, the extension from vector calculus to matrix calculus in Section 5.15, and the differentials of the determinant and inverse in Sections 8.3 and 8.4.

§7. See Section 1.6 for more detailed results.

§8–9. Second differentials are introduced in Chapter 6. The second identification theorem is proved in Section 6.8 and the chain rule for second differentials in Section 6.11.

§11. See Chapter 2 for many more details on the vec operator and the Kronecker product. Theorem 2.2 is restated here as Theorem 18.5.

§12. See Sections 5.15 and 10.2.

§13 and §15. The commutation matrix and the duplication matrix are discussed in Chapter 3.

§16. Many aspects of maximum likelihood estimation are treated in Chapter 15.