Journal of Econometrics 244 (2024) 105862

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A gentle introduction to matrix calculus

Jan R. Magnus

Department of Econometrics & Data Science, Vrije Universiteit Amsterdam, and Tinbergen Institute, Amsterdam, The Netherlands

ARTICLE INFO

JEL classification:
C02
C60
C61

Keywords:
Differential calculus
Multivariate
Optimization
Vec operator
Commutation matrix

ABSTRACT

Matrix calculus is an important tool when we wish to optimize functions involving matrices or perform sensitivity analyses. This tutorial is designed to make matrix calculus more accessible to graduate students and young researchers. It contains the theory that would suffice in most applications, many fully worked-out exercises and examples, and presents some of the 'tacit knowledge' that is prevalent in this field.

1. Introduction

The purpose of this paper is to make matrix calculus more accessible to graduate students and young researchers. It is written as a stand-alone tutorial with minimal references to underlying results, so that it can be used in an advanced undergraduate course and as an aid to graduate and Ph.D. students in one of their courses or projects.

Matrix calculus is much used in economics, statistics, mathematics, psychology, control theory, engineering, and elsewhere, and it is required whenever we wish to optimize functions involving matrices or perform sensitivity analyses. Heinz Neudecker and I started to work on our monograph *Matrix Differential Calculus with Applications in Statistics and Econometrics* in the early 1980s, and the book came out in 1988. It went through several revisions and new editions, and I regard the 2019 third edition, which I wrote after Heinz' death, as the definitive text.

Not everyone is willing to struggle through a 450-page monograph, and it is for those people that the current tutorial may be of some use. It contains the essence of matrix calculus, leaving out the subtleties that are not used in most applications, such as how to differentiate eigenvalues or Moore–Penrose inverses. Obviously, the tutorial provides less information than the monograph, but in two directions it provides more. First, the monograph contains many exercises, but there exists no answer book, so there are no model answers. The current tutorial also contains many exercises, but they all come with fully worked-out solutions. Second, while the emphasis in the monograph is primarily on the theory, the emphasis in this tutorial is primarily on the practice, and I have included suggestions on shortcuts, warnings for pitfalls, and certain 'tricks' that are useful to know when doing matrix calculus in practice.

Matrix calculus rests on two pillars and it requires six tools. The tools are presented in Section 2 and consist of some basic results on the trace and linear and quadratic functions, some less basic results on the Kronecker product and the vec operator, and finally some more advanced results on the commutation and duplication matrix. Almost all these results will be proved using only elementary mathematics.

The two pillars are discussed in Sections 3 and 4. The first pillar is the definition of a matrix derivative, which I present in Section 3. Suppose that we have a matrix function, say $F(X) = X^{-1}$. The derivative $DF(X)$ will be a matrix, but how are the

E-mail address: jan@janmagnus.nl.

https://doi.org/10.1016/j.jeconom.2024.105862
Received 21 April 2024; Received in revised form 23 July 2024; Accepted 8 September 2024
Available online 19 September 2024

elements in this matrix to be organized? This seemingly simple question was not fully resolved until the mid-1980s. I will show that there exists only one correct definition of a matrix derivative.

Section 4 introduces the concept of a differential, the second pillar on which matrix calculus rests. The key advantage of the differential over the more common derivative is the following. Consider an $m\times 1$ vector function f, for example $f(x) = Ax$ where A is an $m\times n$ matrix of constants. Then the derivative $Df(x)$ is an $m\times n$ matrix, in this case the matrix A. But the differential $df(x)$ remains an $m \times 1$ vector. The differential $df(x)$ has the same dimension as f, irrespective of the dimension of the vector x. The advantage is even larger for matrices. The differential $dF(X)$ 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 will be demonstrated through many examples.

I discuss optimization in Section 5 and this leads to our first example (least squares) in Section 6, both with and without constraints. In Section 7, I introduce matrix calculus, showing how the previous results can be straightforwardly generalized from vector calculus to matrix calculus, at least if one employs the correct definition of matrix derivative. Section 8 contains exercises, and Sections 9–11 are devoted to the second differential with associated exercises in Section 12. Sections 13–15 contain three further (more advanced) examples on how to apply matrix calculus. In the concluding Section 16, I provide some hints from my own experience in an attempt to transform tacit (unwritten) knowledge into explicit (written) knowledge.

The following notation is used. Lower-case symbols (a, x) denote scalars or vectors, upper-case symbols (A, X) denote matrices. Thus, f denotes a scalar or vector function, and F a matrix function. I write A', A^{-1}, $\operatorname{tr} A$, $|A|$ for the transpose, inverse, trace, and determinant of A. All functions and variables are real. Parentheses are used sparingly. I write dX, $\operatorname{tr} X$, and $\operatorname{vec} X$ without parentheses, and also dXY, $\operatorname{tr} XY$, and $\operatorname{vec} XY$ instead of $d(XY)$, $\operatorname{tr}(XY)$, and $\operatorname{vec}(XY)$. However, I write $\operatorname{vech}(X)$ with parentheses for historical reasons.

2. The tools

We present six tools that are indispensable in matrix calculus.

2.1. The trace

An important function of a square matrix $A = (a_{ij})$ is the *trace*, defined as the sum of its diagonal elements: $\operatorname{tr} A = \sum_i a_{ii}$. Since a matrix and its transpose contain the same diagonal elements, we have

$$\operatorname{tr} A' = \operatorname{tr} A. \tag{1}$$

Somewhat less trivial is

Proposition 1. *For any two matrices A and B of the same order,*

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

Proof. This follows because

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

The proposition implies that $\operatorname{tr} ABC = \operatorname{tr} CAB = \operatorname{tr} BCA$, because the proposition allows *cyclical* permutations of the matrices. But the proposition does not imply that $\operatorname{tr} ABC = \operatorname{tr} ACB$, which is not cyclical.

2.2. Linear and quadratic forms

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

Things are different with a *quadratic* form, that 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 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.

Proposition 2. *We have*

(i) $Ax = 0$ *for every* $x \iff A = 0$,
(ii) $x'Ax = 0$ *for every* $x \iff A$ *is skew-symmetric,*
(iii) $x'Ax = 0$ *for every* x *and* $A = A' \iff A = 0$.

Proof. We shall prove only one direction of these equivalences, the other being trivial. Let e_i denote the ith elementary vector, that is the vector with one in the ith position and zeros elsewhere. To prove (i), let $x = e_i$. Then, $Ae_i = 0$ for every i, that is, every column of A is zero, and hence $A = 0$. To prove (ii), let $x = e_i + e_j$. Then,

$$a_{ii} + a_{ij} + a_{ji} + a_{jj} = 0,$$

implying in particular, by setting $i = j$, that $a_{ii} = a_{jj} = 0$. Hence, $a_{ij} + a_{ji} = 0$, as requested. To prove (iii), note that if A is symmetric and skew-symmetric, then $A = -A$, implying that A must be the null matrix. □

Proposition 2(i) has important practical consequences. Suppose we wish to prove that $A = B$. Then we could try and prove that $a_{ij} = b_{ij}$ for all i and j. But it is often easier to show that $Ax = Bx$ for arbitrary x. Instead of showing directly that $A = B$, we show that A and B have the same effect when working on an arbitrary vector x.

Proposition 2(ii) also has important consequences. It tells us that if $x'Ax = x'Bx$ for every x, then it is not necessarily true that $A = B$, but it is true that $A + A' = B + B'$. If A is symmetric and $x'Ax = x'Bx$ for every x, then it still does not follow that $A = B$, but it does follow that $A = (B + B')/2$. This will be important when we discuss the second identification theorem in Proposition 13 and Eq. (42).

2.3. 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

$$\begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix} \tag{2}$$

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.

Proposition 3. *The following three properties show that the Kronecker product is in fact a product:*

$$\begin{aligned} (A \otimes B) \otimes C &= A \otimes (B \otimes C), \\ (A + B) \otimes (C + D) &= A \otimes C + A \otimes D + B \otimes C + B \otimes D, \\ (A \otimes B)(C \otimes D) &= AC \otimes BD, \end{aligned}$$

where it is assumed that $A + B$ *and* $C + D$ *are defined in the second equality, and that* AC *and* BD *are defined in the third equality.*

Proof. These equalities are easiest proved by direct application of the definition, which takes quite a bit of space and is therefore omitted. □

Exercise 1. For any two column vectors a and b (not necessarily of the same order), we have

$$a \otimes b' = ab' = b' \otimes a.$$

Solution. This follows, in essence, from the fact that a vector x can be written as $x \otimes 1$ and also as $1 \otimes x$. Define elementary vectors e_i $(m \times 1)$ and u_j $(n \times 1)$, where m and n are the orders of a and b, respectively. Then,

$$\begin{aligned} e_i'(a \otimes b')u_j &= (e_i' \otimes 1)(a \otimes b')(1 \otimes u_j) = (e_i'a) \otimes (b'u_j) = a_i b_j = (ab')_{ij}, \\ e_i'(b' \otimes a)u_j &= (1 \otimes e_i')(b' \otimes a)(u_j \otimes 1) = (b'u_j) \otimes (e_i'a) = b_j a_i = (ab')_{ij}, \end{aligned}$$

using Proposition 3. ‖

The transpose, trace, and inverse of the Kronecker product are given in the next proposition.

Proposition 4. *We have*

$$\begin{aligned} (A \otimes B)' &= A' \otimes B', \\ \operatorname{tr}(A \otimes B) &= (\operatorname{tr} A)(\operatorname{tr} B), \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1}, \end{aligned}$$

where A *and* B *are assumed to be square in the second equality and nonsingular in the third equality.*

Proof. The first two results follow again by direct application of the definition. The third result follows from Proposition 3 by writing

$$(A \otimes B)(A^{-1} \otimes B^{-1}) = (AA^{-1}) \otimes (BB^{-1}) = I_m \otimes I_n = I_{mn},$$

where A has order $m \times m$, and B has order $n \times n$. □

2.4. The vec operator

Consider an $m \times n$ matrix A. This matrix has n columns, say $a_1, \ldots, a_n$. We can transform the matrix into a vector by defining the $mn \times 1$ vector $\operatorname{vec} A$ as the vector which stacks the columns of A one underneath the other:

$$\operatorname{vec} A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}. \tag{3}$$

The vec operator is not the only operator that transforms a matrix into a vector. We could also take the rows of A and stack them one underneath the other, but the vec transformation is now the most common.

Exercise 2. If

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

then what is $\operatorname{vec} A$? And what is $\operatorname{vec} A'$?
Solution. We have $\operatorname{vec} A = (1,4,2,5,3,6)'$ and $\operatorname{vec} A' = (1,2,3,4,5,6)'$. ‖

The vec operator and the trace are related through the following result.

Proposition 5. *If A and B are matrices of the same order, then*

$$(\operatorname{vec} A)'(\operatorname{vec} B) = \operatorname{tr} A'B.$$

Proof. The product $(\operatorname{vec} A)'(\operatorname{vec} B)$ multiplies the corresponding elements in A and B and adds them together, so that $(\operatorname{vec} A)'(\operatorname{vec} B) = \sum_{ij} a_{ij} b_{ij}$. But this is precisely the expression for $\operatorname{tr} A'B$ by the proof of Proposition 1. □

Exercise 3. If a and b are vectors of arbitrary order, show that

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

Solution. The jth column of ab' is given by $b_j a$. ‖

The previous exercise relates the Kronecker product and the vec operator. Its generalization is an important and frequently used tool.

Proposition 6. *For any matrices A, B, and C for which the product ABC is defined, we have*

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

Proof. Using Exercise 3, we see that

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

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, the result follows. □

2.5. 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}. Thus,

$$K_{mn} \operatorname{vec} A = \operatorname{vec} A'. \tag{4}$$

If $m = n$, we write K_n instead of K_{nn}. The commutation matrix is a square matrix with exactly one entry of 1 in each row and each column with all other entries 0. (Such matrices are called permutation matrices.)

Exercise 4. Show that $\operatorname{tr} K_n = n$.
Solution. If A is a square matrix of order $n \times n$, then its ith diagonal element is placed in the same position in $\operatorname{vec} A$ as in $\operatorname{vec} A'$. Hence, the ith diagonal block of K_n has 1 in the iith position and zeros elsewhere. Summing up the diagonal elements of K_n thus adds 1 for each block, that is, n in total. ‖

Proposition 7. *The commutation matrix satisfies*

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

Proof. All permutation matrices are orthogonal, hence $K'_{mn} = K_{mn}^{-1}$. Also, premultiplying (4) by K_{nm} gives $K_{nm}K_{mn} \operatorname{vec} A = \operatorname{vec} A$, which shows that $K_{nm}K_{mn} = I_{mn}$. □

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

Proposition 8.

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

for any $m \times n$ matrix A and $p \times q$ matrix B.

Proof. 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, in the spirit of Proposition 2(i). Thus, let X be an arbitrary $q \times n$ matrix. Then, by repeated application of (4) and Proposition 6,

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

Since X is arbitrary, the result follows. □

Exercise 5. Let $N_n = (I_{n^2} + K_n)/2$. Show that N_n is symmetric idempotent and that $N_n(A \otimes A) = (A \otimes A)N_n$ for every $n \times n$ matrix A.
Solution. Since K_n is symmetric and orthogonal by Proposition 7, we have $K_n^2 = I_{n^2}$ and hence $N_n = N'_n = N_n^2$. Also, using Proposition 8, $K_n(A \otimes A) = (A \otimes A)K_n$ and hence $N_n(A \otimes A) = (A \otimes A)N_n$. ‖

2.6. The duplication matrix

Many matrices in statistics and econometrics are symmetric. 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 $\operatorname{vech}(A)$ (the 'vec-half' operator) denotes the $\frac{1}{2}n(n+1) \times 1$ vector that is obtained from $\operatorname{vec} A$ by eliminating all elements of A above the diagonal. For example, when $n = 3$, crossing out the elements above the diagonal,

$$\begin{pmatrix} a_{11} & \cancel{a_{12}} & \cancel{a_{13}} \\ a_{21} & a_{22} & \cancel{a_{23}} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

and vectorizing the remaining elements, we obtain

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

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

$$D_n \operatorname{vech}(A) = \operatorname{vec} A \qquad (A = A'). \tag{5}$$

The matrix D_n has full column rank $\frac{1}{2}n(n+1)$, so that D'_nD_n is nonsingular. This implies that $\operatorname{vech}(A)$ can be uniquely solved from (5), and we have

$$\operatorname{vech}(A) = (D'_nD_n)^{-1}D'_n \operatorname{vec} A \qquad (A = A'). \tag{6}$$

Proposition 9. *The duplication matrix is connected to the commutation matrix by*

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

Proof. Let X be a symmetric $n \times n$ matrix. Then,

$$K_n D_n \operatorname{vech}(X) = K_n \operatorname{vec} X = \operatorname{vec} X = D_n \operatorname{vech}(X).$$

The symmetry of X does not restrict $\operatorname{vech}(X)$, which is therefore arbitrary. Hence, the first result follows. To prove the second result, let

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

Both M_n and N_n are symmetric idempotent, the latter by Exercise 5. Since $K_n D_n = D_n$, we have $M_n N_n = N_n M_n = M_n$, so that Δ_n is also symmetric idempotent. Now,

$$\operatorname{tr} M_n = \operatorname{tr} I_{n(n+1)/2} = n(n+1)/2 = (n^2+n)/2 = \operatorname{tr} N_n,$$

since $\operatorname{tr} K_n = n$ (Exercise 4). This gives $r(\Delta_n) = \operatorname{tr} \Delta_n = 0$, and hence $\Delta_n = 0$. □

Much of the interest in the duplication matrix is due to the importance of the matrix $D_n'(A \otimes A)D_n$, caused by the fact that, for symmetric X, the scalar expression $\operatorname{tr} AXA'X$ occurs frequently in statistics and econometrics; see Exercise 28 and Section 13.

Proposition 10. *Let A be an $n \times n$ matrix, not necessarily symmetric. The matrix $D_n'(A \otimes A)D_n$ satisfies the following properties:*

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

where A is assumed to be nonsingular in the second equality.

Proof. The first result follows from Exercise 5 and $N_n D_n = D_n$. To prove the second result we write

$$\begin{aligned} &D_n'(A \otimes A)D_n(D_n'D_n)^{-1}D_n'(A^{-1} \otimes A^{-1})D_n(D_n'D_n)^{-1} \\ &\quad = D_n'(A \otimes A)(A^{-1} \otimes A^{-1})D_n(D_n'D_n)^{-1} \\ &\quad = D_n'D_n(D_n'D_n)^{-1} = I_{n(n+1)/2}. \end{aligned}$$

The third result is more difficult to prove. It follows from

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

and the fact that $|D_n'D_n| = 2^{\frac{1}{2}n(n-1)}$ and the eigenvalues of the matrix $(D_n'D_n)^{-1}D_n'(A \otimes A)D_n$ are $\lambda_i\lambda_j$ $(1 \le j \le i \le n)$ where $\{\lambda_i\}$ denote the eigenvalues of A; see Magnus (1988, Theorems 4.4 and 4.10). □

3. The definition of matrix derivative

Matrix calculus uses six tools (discussed in the previous section) and it rests on two pillars: the correct definition of a matrix derivative and the concept of a differential. In the current section I discuss how to define a matrix derivative; in the next section I introduce differentials.

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? Let us 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

$$\frac{\partial f(x)}{\partial x'} = A. \tag{7}$$

The notation $\partial f(x)/\partial x'$ is just notation, nothing else. We sometimes write the derivative as $Df(x)$ or as $f'(x)$, but we only use the latter notation if it does not 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

$$\frac{\partial f(x)}{\partial x'} = \begin{pmatrix} \partial f_1(x)/\partial x_1 & \partial f_1(x)/\partial x_2 & \dots & \partial f_1(x)/\partial x_n \\ \partial f_2(x)/\partial x_1 & \partial f_2(x)/\partial x_2 & \dots & \partial f_2(x)/\partial x_n \\ \vdots & \vdots & & \vdots \\ \partial f_m(x)/\partial x_1 & \partial f_m(x)/\partial x_2 & \dots & \partial f_m(x)/\partial x_n \end{pmatrix}. \tag{8}$$

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 (8) 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, 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 essence of the concept derivative must be preserved when we generalize from vector functions $f(x)$ to matrix functions $F(X)$. While it is tempting to keep the structure of the matrices X and F intact — and this line of thought was the original idea about matrix derivatives at the end of the 1960s — it was not until the early 1980s that we realized that this is the wrong generalization, since it does not maintain the concept of derivative; see Magnus (2010) for further discussion on the concept of matrix derivative and Magnus (2024) for a sketch of the historical development.

We now understand that when we have a collection of functions and a collection of variables, which may be organized in matrices (such as F and X) or otherwise, then what is required is a *one*-dimensional ordering of the functions and of the variables, for which we choose vec F and vec X, because these are now the most common vectorizations. In fact, which vectorization we choose is irrelevant, we may even use a different vectorization for F and for X, although this would be confusing and is not recommended.

However, once we have chosen our vectorization, we must stick to it, so that the derivative

$$\frac{\partial \operatorname{vec} F(X)}{\partial (\operatorname{vec} X)'} \tag{9}$$

has the essential property that each row 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 the only correct definition of a matrix derivative.

4. The differential

4.1. Definition, notation, and rules

If f is a scalar function of one variable, such as $f(x) = x^2$, then we define the differential $df(x)$ at a point x as the linear function (of u)

$$df(x)(u) = f'(x)u, \tag{10}$$

where x is a point where the derivative $f'(x)$ exists and u is an arbitrary point in $\mathbb{R}$. The differential is closely related to the derivative, but the two concepts are not the same: the differential is a geometric concept, while the derivative is an algebraic concept, representing the differential by some number (the slope or the Jacobian).

Exercise 6. Show that

$$\begin{aligned} f(x) = x &\implies df(x)(u) = u, \\ f(x) = x^2 &\implies df(x)(u) = 2xu, \\ f(x) = \sin(x) &\implies df(x)(u) = \cos(x)u, \\ f(x) = \sin^2(x) &\implies df(x)(u) = 2\sin(x)\cos(x)u. \end{aligned}$$

Solution. These results follow directly from the definition in (10), where in the last case we apply the chain rule:

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

Exercise 6 shows that the differential associated with the identity function $f(x) = x$ is given by $dx(u) = u$. Hence, $df(x)(u) = f'(x)\,dx(u)$, or for short,

$$df(x) = f'(x)\,dx, \tag{11}$$

which is the notation we shall use hereafter.

Exercise 7. Show that

$$\begin{aligned} dx^2 &= 2x\,dx, \\ d\sin(x) &= \cos(x)\,dx, \\ d\sin^2(x) &= 2\sin(x)\cos(x)\,dx. \end{aligned}$$

Solution. This follows immediately from the previous exercise by replacing u with dx. $\|$

The differential is an operator, in fact a linear operator, and it obeys the following simple rules:

$$\begin{aligned} da &= 0, \qquad (a \text{ constant}), \\ d(ax) &= a\,dx \qquad (a \text{ constant}), \\ d(f(x) + g(x)) &= df(x) + dg(x), \\ d(f(x)g(x)) &= (df(x))\,g(x) + f(x)\,dg(x). \end{aligned} \tag{12}$$

4.2. Cauchy's rule of invariance

The chain rule, well-known for derivatives, also applies to differentials and is then called *Cauchy's rule of invariance*. Cauchy's rule states that taking differentials of functions preserves composition. Formally, if $z = h(x) = f(g(x))$ is the composition of the functions $y = g(x)$ and $z = f(y)$, then $dy = g'(x)\,dx$ and $dz = f'(y)\,dy$, so that

$$dz = f'(y)\,dy = f'(y)g'(x)\,dx = f'(g(x))\,g'(x)\,dx. \tag{13}$$

Exercise 8. Use Cauchy's rule of invariance to find the differentials of the functions $\sin^2(x)$, e^{x^2}, and $e^{\sin(x^2)}$.
Solution. We have

$$\begin{aligned} d\sin^2(x) &= 2\sin(x)\,d\sin(x) = 2\sin(x)\cos(x)\,dx, \\ de^{x^2} &= e^{x^2}dx^2 = 2e^{x^2}x\,dx, \\ de^{\sin(x^2)} &= e^{\sin(x^2)}\,d\sin(x^2) = e^{\sin(x^2)}\cos(x^2)\,dx^2 \\ &= 2x\,e^{\sin(x^2)}\cos(x^2)\,dx. \end{aligned}$$

Notice the subtle difference between the derivation of $d\sin^2(x)$ here, compared with Exercises 6 and 7. In Exercise 6 we used the chain rule, while here we use Cauchy's rule of invariance. The latter is simpler. ‖

Cauchy's rule, like the chain rule, is a key instrument in differential calculus. Suppose we realize that x in the function $\sin^2(x)$ depends on t, say $x = t^2$. Then, we do not need to compute the differential of $\sin^2(t^2)$ all over again. Instead, we write simply

$$d\sin^2(t^2) = 2\sin(t^2)\cos(t^2)\,dt^2 = 4t\sin(t^2)\cos(t^2)\,dt.$$

Cauchy's rule thus allows us to apply the rules of calculus sequentially, one after another.

If we know the derivative then we can compute the differential. For our purposes we need the opposite: to compute the derivative from the differential. This is possible, because

$$df(x) = \alpha(x)\,dx \iff f'(x) = \alpha(x), \tag{14}$$

where α may depend on x, but not on dx. Eq. (14) is a special case of the *first identification theorem* (Proposition 11). It shows that we can *identify* the derivative from the differential (and vice versa), and it shows that the concept differential is equivalent to the familiar concept derivative.

4.3. Vector functions

So far we have only concerned ourselves with scalar functions of one variable, 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 may be true for scalar functions of one variable, but it is not true for vector functions of several variables, to which we now turn.

Thus, let f now denote an $m\times 1$ vector function of an $n\times 1$ vector x. The derivative $Df(x)$ is an $m\times n$ matrix, but the differential $df(x)$ has the same dimension as f, namely $m\times 1$. The practical advantage of working with differentials is therefore that it preserves the order of the function. This simple fact will be of great practical importance, especially when we generalize from vectors to matrices.

The differential of a vector function f is defined as

$$df(x)(u) = (Df(x))\,u, \tag{15}$$

generalizing (10), where, as in the one-dimensional case, the identity function $f(x) = x$ gives $dx(u) = u$, which allows us to write

$$df(x) = (Df(x))\,dx, \tag{16}$$

generalizing (11).

Exercise 9. Find the differential of the function $f(x) = x_1^2 x_2$.
Solution. We have

$$df(x) = \frac{\partial f(x)}{\partial x_1}\,dx_1 + \frac{\partial f(x)}{\partial x_2}\,dx_2 = 2x_1x_2\,dx_1 + x_1^2\,dx_2,$$

where we note that both $f(x)$ and $df(x)$ are scalars. ‖

The operating rules (12) concerning scalar functions of one variable generalize straightforwardly to vector functions of several variables, and we have

$$\begin{aligned} &da = 0, \qquad d(x') = (dx)', \qquad d(a'x) = a'dx, \\ &d(x+y) = dx + dy, \qquad d(x'y) = (dx)'y + x'dy, \end{aligned} \tag{17}$$

where x and y are vectors and a is a vector of real constants, all of the same order.

In Exercise 9 we found the differential by first calculating the derivative, but our purpose is to find the derivative by first calculating the differential. This would work like this.

Exercise 10. Find again the differential of the function $f(x) = x_1^2 x_2$ without first calculating the partial derivatives.
Solution. Using the operating rules, we write

$$\begin{aligned} df(x) &= d\left(x_1^2 x_2\right) = (dx_1^2)x_2 + x_1^2(dx_2) = 2x_1 x_2\, dx_1 + x_1^2\, dx_2 \\ &= (2x_1 x_2,\ x_1^2)\begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix}, \end{aligned}$$

from which we find the derivative as $Df(x) = (2x_1 x_2,\ x_1^2)$. ‖

4.4. First identification theorem

By identifying the derivative from the differential, we have generalized the identification result (14) from scalar functions of one variable to vector functions of several variables. This one-to-one relationship is formally stated as follows.

Proposition 11 (*First Identification Theorem*)**.**

$$df(x) = A(x)\,dx \iff Df(x) = A(x).$$

Let us apply Proposition 11 to linear and quadratic functions, the most frequently used functions in statistics and econometrics.

Exercise 11. Find the derivative of the linear function $a'x$, where a is a vector of constants, and of the quadratic function $x'Ax$, where A is a square matrix of constants.
Solution. Writing out all the steps (which one would not normally do), we have

$$\begin{aligned} d(a'x) &= (da)'x + a'(dx) = a'\,dx, \\ d(x'Ax) &= (dx)'Ax + x'(dA)x + x'A(dx) \\ &= x'A'\,dx + x'A\,dx = x'(A + A')\,dx, \end{aligned}$$

where we used the fact that the differential of a constant is 0, and also the fact that the transpose of a scalar is the same scalar, so that $(dx)'Ax = x'A'\,dx$. Hence, the derivatives are

$$\frac{\partial a'x}{\partial x'} = a', \qquad \frac{\partial x'Ax}{\partial x'} = x'(A + A'),$$

and in the special case where A is symmetric, the derivative of $x'Ax$ is $2x'A$. In either case, the derivative is a *row* vector, not a column vector. ‖

4.5. Cauchy invariance for vector functions

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

$$\frac{\partial z}{\partial x'} = \frac{\partial z}{\partial y'}\,\frac{\partial y}{\partial x'}. \tag{18}$$

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

Proposition 12 (*Cauchy Invariance*)**.** *Let $z = f(y)$ and $y = g(x)$, so that $z = f(g(x))$. Then,*

$$dz = A(y)B(x)\,dx,$$

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

$$dz = A(y)\,dy, \qquad dy = B(x)\,dx.$$

In practice, we avoid introducing a new variable z or a new function $h(x) = f(g(x))$, and we write $f(y)$ when we think of z as a function of y and we write $f(x)$ when we think of z as a function of x. With this convention, Proposition 12 can be reformulated as

$$df(x) = A(y)B(x)\,dx, \tag{19}$$

where $df(y) = A(y)\,dy$ and $dy = dg(x) = B(x)\,dx$. Such abuses of notation are very common in mathematics when one has to balance pedantic correctness and readability. In fact, the chain rule (18) is typically written as $\partial f(x)/\partial x' = (\partial f(y)/\partial y')(\partial y/\partial x')$, without anybody complaining about it.

Exercise 12. Find the derivative of

$$f(x) = \begin{pmatrix} x_1^2 - x_2^2 \\ x_1 x_2 x_3 \end{pmatrix}, \qquad x = (x_1, x_2, x_3)'.$$

Solution. The differential is

$$\begin{aligned} df(x) &= \begin{pmatrix} d(x_1^2) - d(x_2^2) \\ d(x_1 x_2 x_3) \end{pmatrix} = \begin{pmatrix} 2x_1 dx_1 - 2x_2\, dx_2 \\ (dx_1)x_2 x_3 + x_1 (dx_2) x_3 + x_1 x_2\, dx_3 \end{pmatrix} \\ &= \begin{pmatrix} 2x_1 & -2x_2 & 0 \\ x_2 x_3 & x_1 x_3 & x_1 x_2 \end{pmatrix} \begin{pmatrix} dx_1 \\ dx_2 \\ dx_3 \end{pmatrix}, \end{aligned}$$

which identifies the derivative as

$$\frac{\partial f(x)}{\partial x'} = \begin{pmatrix} 2x_1 & -2x_2 & 0 \\ x_2 x_3 & x_1 x_3 & x_1 x_2 \end{pmatrix}. \quad \|$$

Exercise 13. Let

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

Find the derivative of f with respect to $x = (x_1, x_2)'$.
Solution. Using the notational convention (19), we write

$$\begin{aligned} df(y) &= (de^{y_1}) \sin(y_2) + e^{y_1}\, d \sin(y_2) \\ &= e^{y_1} \sin(y_2)\, dy_1 + e^{y_1} \cos(y_2)\, dy_2 = a(y)' dy, \end{aligned}$$

where

$$a(y) = e^{y_1} \begin{pmatrix} \sin(y_2) \\ \cos(y_2) \end{pmatrix}, \qquad dy = \begin{pmatrix} dy_1 \\ dy_2 \end{pmatrix}.$$

Also,

$$dy = \begin{pmatrix} x_2^2 & 2x_1 x_2 \\ 2x_1 x_2 & x_1^2 \end{pmatrix} \begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix} = B(x)\, dx.$$

Hence,

$$df(x) = a(y)'\, dy = a(y)' B(x)\, dx = c_1\, dx_1 + c_2\, dx_2,$$

where

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

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

5. Optimization

Let $f(x)$ be a scalar function that we wish to optimize with respect to an $n \times 1$ vector x. We achieve this by first computing the differential $df(x) = a(x)'\, dx$, and then setting $a(x) = 0$.

Exercise 14. Minimize the function

$$f(x) = \frac{1}{2} x' A x - b' x,$$

where the matrix A is positive definite.
Solution. The differential is

$$df(x) = x' A\, dx - b'\, dx = (Ax - b)'\, dx,$$

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

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

[1] A (semi)definite matrix *must* be symmetric. For example, $A = \begin{pmatrix} 1 & 3 \\ -6 & 9 \end{pmatrix}$ satisfies $x' A x > 0$ for all nonzero x, but its eigenvalues are not real, let alone positive. This matrix is *not* positive definite.

Since A is positive definite, $y'Ay$ has a minimum at $y = 0$ and hence $f(x)$ has a minimum at $x = \hat{x}$. Alternatively, we note that $f(x)$ is strictly convex and use the fact that any (strictly) convex function attains a (strict) absolute minimum. ‖

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

$$\mathcal{L}(x) = f(x) - \lambda g(x), \tag{20}$$

where λ is the Lagrange multiplier. Then we obtain the differential of $\mathcal{L}(x)$ with respect to x,

$$d\mathcal{L}(x) = df(x) - \lambda dg(x), \tag{21}$$

and set it equal to zero. The equations

$$\frac{\partial f(x)}{\partial x'} = \lambda \frac{\partial g(x)}{\partial x'}, \qquad g(x) = 0 \tag{22}$$

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

Exercise 15. Let A be positive definite. Minimize the function

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

subject to the constraint $c'x = 0$.
Solution. We write the Lagrangian as

$$\mathcal{L}(x) = \frac{1}{2}x'Ax - b'x - \lambda c'x.$$

The differential of $\mathcal{L}(x)$ is

$$d\mathcal{L}(x) = (Ax - b)'\,dx - \lambda c'\,dx = (Ax - b - \lambda c)'\,dx,$$

from which we obtain the first-order conditions $Ax - b - \lambda c = 0$ and $c'x = 0$. This gives $x = A^{-1}(b + \lambda c)$ and hence $0 = c'x = c'A^{-1}(b + \lambda c)$. We then solve $\tilde{\lambda} = -c'A^{-1}b/c'A^{-1}c$, so that the constrained optimum is achieved at

$$\tilde{x} = A^{-1}(b + \tilde{\lambda}c) = A^{-1}b - \frac{c'A^{-1}b}{c'A^{-1}c}A^{-1}c.$$

Alternatively, we can write the first-order conditions as

$$\begin{pmatrix} A & c \\ c' & 0 \end{pmatrix}\begin{pmatrix} x \\ -\lambda \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

and solve for x and $-\lambda$ by inverting the matrix on the left-hand side. Notice that by expressing the equation in x and $-\lambda$ (rather than x and λ) I achieve symmetry of the matrix on the left-hand side. Symmetric matrices are easier, both theoretically and computationally, than non-symmetric matrices, so, if possible, we work with symmetric matrices.

Either way, since $f(x)$ is linear-quadratic (hence strictly convex) and the constraint is linear, $f(x)$ attains an absolute minimum at $\tilde{x}$ under the constraint. ‖

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

$$\mathcal{L}(x) = f(x) - l'g(x), \tag{23}$$

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

$$d\mathcal{L}(x) = df(x) - l'dg(x) \tag{24}$$

and set it equal to zero. The equations

$$\frac{\partial f(x)}{\partial x'} = l'\frac{\partial g(x)}{\partial x'}, \qquad g(x) = 0 \tag{25}$$

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

Exercise 16. Show that optimizing a function subject to constraints is not equivalent to optimizing the Lagrangian function.
Solution. A simple counterexample is provided by the optimization problem: $\max(xy)$ under the constraint $x + y = 2$. The solution is $x = y = 1$, but the Lagrangian $\mathcal{L}(x, y) = xy - \lambda(x + y - 2)$ has a saddle-point at $(1, 1)$. ‖

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 often occurring situation where $\mathcal{L}(x)$ is (strictly) convex, as in Exercises 14 and 15, $f(x)$ attains a (strict) absolute

minimum at the solution $\hat{x}$ under the constraint $g(x) = 0$. Notice that the Lagrangian $\mathcal{L}(x, y)$ in Exercise 16 is not convex, because $\partial\mathcal{L}(x, y)/\partial x = y - \lambda$ and $\partial\mathcal{L}(x, y)/\partial y = x - \lambda$, so that the Hessian takes the form

$$H(x, y) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which is not positive semidefinite.

6. Example 1: Least squares

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

$$f(x) = \frac{e'e}{2} = \frac{(b - Ax)'(b - Ax)}{2}, \tag{26}$$

where we note that we write $e'e/2$ rather than $e'e$. This makes no difference, since any x 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 $f(x)$ in (26) gives

$$df(x) = e'\,de = e'\,d(b - Ax) = -e'A\,dx.$$

Hence, the optimum is obtained when $A'e = 0$, that is, when $A'Ax = A'b$, from which we obtain

$$\hat{x} = (A'A)^{-1}A'b, \tag{27}$$

the least-squares solution.

If there are constraints on x, say $Rx = r$, then we need to solve

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad Rx = r. \end{aligned} \tag{28}$$

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

$$\mathcal{L}(x) = (b - Ax)'(b - Ax)/2 - l'(Rx - r),$$

where l is a vector of Lagrange multipliers.

Exercise 17. Show that the matrix $V = R(A'A)^{-1}R'$ is nonsingular if and only if the m rows of R are linearly independent.
Solution. Let $B = R(A'A)^{-1/2}$, so that $r(B) = r(R)$. Since $V = BB'$, it follows that $r(V) = r(B)$, and hence that $r(V) = r(R)$. Since V is an $m \times m$ matrix, it is nonsingular if and only if $r(V) = m$, that is, if and only if $r(R) = m$. ‖

We now write the differential of $\mathcal{L}(x)$ as

$$\begin{aligned} d\mathcal{L}(x) &= d(b - Ax)'(b - Ax)/2 - l'\,d(Rx - r) \\ &= (b - Ax)'\,d(b - Ax) - l'R\,dx \\ &= -(b - Ax)'A\,dx - l'R\,dx. \end{aligned}$$

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

$$(b - A\tilde{x})'A + \tilde{l}'R = 0, \qquad R\tilde{x} = r,$$

or, written differently,

$$A'A\tilde{x} - A'b = R'\tilde{l}, \qquad R\tilde{x} = r.$$

We do not know $\tilde{x}$ but we know $R\tilde{x}$. Hence, we premultiply by $R(A'A)^{-1}$. Letting $\hat{x} = (A'A)^{-1}A'b$ as in (27), this gives

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

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

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

and hence for x:

$$\tilde{x} = \hat{x} + (A'A)^{-1}R'\tilde{l} = \hat{x} + (A'A)^{-1}R'\left(R(A'A)^{-1}R'\right)^{-1}(r - R\hat{x}). \tag{29}$$

Since the constraint is linear and the function $f(x)$ is linear-quadratic, it follows that the solution $\tilde{x}$ indeed minimizes $f(x) = e'e/2$ under the constraint $Rx = r$.

7. Matrix calculus

We have moved from scalar calculus to vector calculus, now we move from vector calculus to matrix calculus. The rules for vector differentials in Section 4 carry over to matrix differentials. Let A be a matrix of constants and let α be a scalar. Then, for any X,

$$dA = 0, \qquad d(\alpha X) = \alpha\, dX, \qquad d(X') = (dX)',$$

and, for square X,

$$d\ \mathrm{tr}\ X = \mathrm{tr}\ dX.$$

If X and Y are of the same order, then

$$d(X + Y) = dX + dY,$$

and, if the matrix product XY is defined,

$$d(XY) = (dX)Y + X dY.$$

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

$$d|X| = |X|\,\mathrm{tr}\,X^{-1}\,dX, \tag{30}$$

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

$$d\log|X| = \frac{d|X|}{|X|} = \mathrm{tr}\,X^{-1}\,dX. \tag{31}$$

The proof of (30) is a little tricky and can be found in *Matrix Differential Calculus*, Section 8.3, where two proofs are provided. The differential of the inverse is, for nonsingular X,

$$dX^{-1} = -X^{-1}(dX)X^{-1}. \tag{32}$$

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

$$(dX^{-1})X + X^{-1}dX = 0$$

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

If we use the correct definition of matrix derivative (but only then), all results from vector calculus carry over to matrix calculus. In particular, the first equivalence theorem for vector functions (Proposition 11) now becomes

$$d\,\mathrm{vec}\,F(X) = A(X)d\,\mathrm{vec}\,X \iff DF(X) = A(X). \tag{33}$$

The same holds for the chain rule. More precisely, if $Z = F(Y)$ and $Y = G(X)$, so that $Z = F(G(X))$, then

$$dZ = A(Y)B(X)\,dX,$$

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

$$dZ = A(Y)\,dY, \qquad dY = B(X)\,dX,$$

as in Proposition 12.

Constrained optimization, treated for vector functions in Section 5, can easily and elegantly be 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

$$\mathcal{L}(X) = f(X) - \mathrm{tr}\,L'G(X), \tag{34}$$

where we have used the fact, as in the proof of Proposition 1, that

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

If the constraint matrix G is symmetric, we may take the matrix of Lagrange multipliers to be symmetric as well.

Exercise 18. Consider the optimization problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } G(X) = 0. \end{aligned}$$

If $G(X)$ is symmetric for all X, then the Lagrangian function is

$$\mathcal{L}(X) = f(X) - \operatorname{tr} LG(X),$$

where L may be assumed to be symmetric.
Solution. Since G is symmetric, we have

$$\operatorname{tr} L'G = \operatorname{tr} L'G' = \operatorname{tr}(GL)' = \operatorname{tr} GL = \operatorname{tr} LG,$$

so that $\operatorname{tr} L'G = \operatorname{tr} L_*G$, where $L_* = (L + L')/2$ is symmetric. ‖

From the Lagrangian (34) we obtain the differential

$$d\mathcal{L}(X) = df(X) - \operatorname{tr} L'\, dG(X), \tag{35}$$

and, setting the differential equal to 0, we obtain the first-order conditions

$$\frac{\partial f(X)}{\partial (\operatorname{vec} X)'} = (\operatorname{vec} L)' \frac{\partial \operatorname{vec} G(X)}{\partial (\operatorname{vec} X)'}, \qquad G(X) = 0, \tag{36}$$

using Proposition 5 and the fact that

$$\operatorname{tr} L'dG(X) = (\operatorname{vec} L)'\, d\operatorname{vec} G(X) = (\operatorname{vec} L)' \frac{\partial \operatorname{vec} G(X)}{(\partial \operatorname{vec} X)'}\, d\operatorname{vec} X.$$

8. Four exercises: first derivative

Exercise 19. Obtain the derivative of the scalar function $f(X) = \operatorname{tr} X'AX$.
Solution. The differential is

$$\begin{aligned} df(X) &= d(\operatorname{tr} X'AX) = \operatorname{tr} d(X'AX) \\ &= \operatorname{tr}(dX)'AX + \operatorname{tr} X'A\, dX = \operatorname{tr} X'(A + A')\, dX \\ &= \operatorname{tr} C'dX = (\operatorname{vec} C)'d\operatorname{vec} X, \end{aligned}$$

using Proposition 5 and letting $C = (A + A')X$. Hence the derivative is $Df(X) = (\operatorname{vec} C)'$.

Exercise 20. Obtain the derivative of the scalar function $f(X) = \log|X'X|$, where X has full column rank.
Solution. From the differential

$$\begin{aligned} df(X) &= d\log|X'X| = \operatorname{tr}(X'X)^{-1}d(X'X) \\ &= \operatorname{tr}(X'X)^{-1}(dX)'X + \operatorname{tr}(X'X)^{-1}X'dX = 2\operatorname{tr}(X'X)^{-1}X'dX \\ &= 2\operatorname{tr} C'dX = 2(\operatorname{vec} C)'d\operatorname{vec} X, \end{aligned}$$

where $C = X(X'X)^{-1}$, we obtain $Df(X) = 2(\operatorname{vec} C)'$.

Exercise 21. Let $f_k(X) = \operatorname{tr} X^k$ $(k = 1, 2, \dots)$. Find the derivative.
Solution. We have

$$\begin{aligned} df_k(X) &= \operatorname{tr}(dX)X^{k-1} + \operatorname{tr} X(dX)X^{k-2} + \cdots + \operatorname{tr} X^{k-1}dX \\ &= k\operatorname{tr} X^{k-1}dX = k(\operatorname{vec} X'^{k-1})'d\operatorname{vec} X. \end{aligned}$$

This gives $Df_k(X) = k(\operatorname{vec} X'^{k-1})'$. In particular,

$$Df_1(X) = D\operatorname{tr} X = (\operatorname{vec} I)', \qquad Df_2(X) = D\operatorname{tr} X^2 = 2(\operatorname{vec} X')'.$$

Exercise 22. Find the derivative of the matrix equation $F(X) = AX^{-1}B$, where X is nonsingular.
Solution. Since

$$dF(X) = A(dX^{-1})B = -AX^{-1}(dX)X^{-1}B,$$

we find

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

using Proposition 6, so that the derivative is given by

$$DF(X) = \frac{\partial \operatorname{vec} F(X)}{\partial (\operatorname{vec} X)'} = -(X^{-1}B)' \otimes (AX^{-1}),$$

by the first identification theorem for matrices (33).

9. The second differential

The second differential is simply the differential of the first differential, that is, $d^2 f = d(df)$. Higher-order differentials are similarly defined, but they are seldom needed.

Exercise 23. Let $f(x) = x'Ax$. Show that $d^2 f(x) = (dx)'(A + A')\,dx$.
Solution. We know from Exercise 11 that $df(x) = x'(A + A')\,dx$. Then,

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

since $d^2 x = 0$. ‖

Exercise 24. Why is $d^2 x = 0$?
Solution. If f is a function of x, then dx is short-hand for the differential $dx(u) = u$ associated with the identity function, see Section 4.1. The first derivative of the identity function is 1 (if f is a scalar function of one variable), and the second derivative is 0. Hence $d^2 x = 0$. In simpler words, if f is a function of x and x is the 'endpoint', then $d^2 x = 0$. But if x is only an intermediary variable and in fact $x = x(t)$, then $d^2 x$ is not 0 (unless x is a linear function of t). This relationship is further developed and made more precise in Section 10. ‖

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 component 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 f be a scalar function and let

$$df(x) = a(x)'dx, \qquad da(x) = (Hf(x))\,dx, \tag{37}$$

where

$$a(x)' = \frac{\partial f(x)}{\partial x'}, \qquad Hf(x) = \frac{\partial a(x)}{\partial x'} = \frac{\partial}{\partial x'}\left(\frac{\partial f(x)}{\partial x'}\right)'.$$

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

The Hessian matrix is often written as

$$Hf(x) = \frac{\partial^2 f(x)}{\partial x\,\partial x'}, \tag{38}$$

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

$$\frac{\partial^2 f(x)}{\partial x\,\partial x'} = \frac{\partial}{\partial x'}\left(\frac{\partial f(x)}{\partial x'}\right)'. \tag{39}$$

Given (37) and using the symmetry of $Hf(x)$, we obtain the second differential as

$$d^2 f(x) = (da)'\,dx = (dx)'(Hf(x))\,dx, \tag{40}$$

which shows that the second differential of f is a quadratic form in dx.

Now, suppose that we have obtained $d^2 f(x) = (dx)'B(x)\,dx$. We also know that $d^2 f(x) = (dx)'Hf(x)\,dx$ by the definition of the Hessian. Hence,

$$(dx)'(Hf(x) - B(x))\,dx = 0$$

for all dx, where $Hf(x)$ is symmetric, but $B(x)$ is not necessarily symmetric. Does this imply that $Hf(x) = B(x)$? No, it does not, as we have seen in Section 2.2. It does, however, imply that

$$(Hf(x) - B(x))' + (Hf(x) - B(x)) = 0,$$

and hence that $Hf(x) = (B(x) + B(x)')/2$, using the symmetry of $Hf(x)$. This proves the following result.

Proposition 13 (*Second Identification Theorem*)**.**

$$d^2 f(x) = (dx)'B(x)\,dx \iff Hf(x) = \frac{B(x) + B(x)'}{2}.$$

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.

Exercise 25. Consider again the quadratic function $f(x) = x'Ax$. Find the second differential without writing out the first differential in its final form.
Solution. We can start with $df(x) = x'(A+A')\,dx$, as in Exercise 23, and obtain $d^2f(x) = (dx)'(A+A')\,dx$. The matrix in the quadratic form is already symmetric, so we obtain directly $Hf(x) = A + A'$. Alternatively — and this is often quicker — we differentiate f twice without writing out the first differential in its final form, as follows. From

$$df(x) = (dx)'Ax + x'A\,dx,$$

we obtain

$$d^2f(x) = 2(dx)'A\,dx,$$

which immediately identifies the Hessian matrix as $Hf(x) = A + A'$. ‖

Even with such a simple function as $f(x) = x'Ax$, 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 (39) 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.

10. Chain rule for second differentials

Let us now further analyze the question why sometimes $d^2x = 0$ and sometimes $d^2x \neq 0$, first considered in Exercise 24. If f is a function of x, and x is the argument of interest, then $d^2x = 0$. But if f 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,

$$dz = A(y)\,dy$$

and

$$d^2z = (dA)\,dy + A(y)\,d^2y. \tag{41}$$

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

$$dy = B(x)\,dx, \qquad d^2y = (dB)\,dx.$$

This leads to the following result.

Proposition 14 (*Chain Rule for Second Differentials*). *Let $z = f(y)$ and $y = g(x)$, so that $z = f(g(x))$. Then,*

$$d^2z = (dA)B(x)\,dx + A(y)(dB)\,dx,$$

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

$$dz = A(y)\,dy, \qquad dy = B(x)\,dx.$$

Personally, I usually avoid Proposition 14. Instead, I express d^2z in terms of dy and d^2y, as in (41), and proceed from there. Let me give two examples, one using Proposition 14, the other not using the proposition.

Exercise 26. Let

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

Find the Hessian matrix by using Proposition 14.
Solution. By Proposition 14,

$$d^2f(x) = (da)'B(x)\,dx + a(y)'(dB)\,dx,$$

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

$$da = C(y)\,dy = C(y)B(x)\,dx$$

and

$$dB = (dx_1)D_1(x) + (dx_2)D_2(x).$$

Let us write dx_1 and dx_2 in terms of dx, which can be done by defining $e_1 = (1,0)'$ and $e_2 = (0,1)'$. Then, $dx_1 = e_1'dx$ and $dx_2 = e_2'dx$, and hence

$$\begin{aligned} d^2 f(x) &= (da)'B(x)\,dx + a(y)'(dB)\,dx \\ &= (dx)'B(x)C(y)B(x)\,dx + a(y)'\left((dx_1)D_1(x) + (dx_2)D_2(x)\right)\,dx \\ &= (dx)'B(x)C(y)B(x)\,dx + (dx)'e_1 a(y)'D_1(x)\,dx + (dx)'e_2 a(y)'D_2(x)\,dx \\ &= (dx)'\left[B(x)C(y)B(x) + e_1 a(y)'D_1(x) + e_2 a(y)'D_2(x)\right]\,dx. \end{aligned}$$

Some care is required where to position the scalars $e_1'dx$ and $e_2'dx$ 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, Proposition 13 implies that the Hessian is equal to

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

Exercise 27. Now find the Hessian matrix, not using Proposition 14.
Solution. We have

$$df(y) = (de^{y_1})\sin(y_2) + e^{y_1}d\sin(y_2) = e^{y_1}\sin(y_2)\,dy_1 + e^{y_1}\cos(y_2)\,dy_2,$$

and hence

$$\begin{aligned} d^2 f(y) &= [de^{y_1}]\sin(y_2)\,dy_1 + e^{y_1}[d\sin(y_2)]\,dy_1 + e^{y_1}\sin(y_2)\,d^2 y_1 \\ &\quad + [de^{y_1}]\cos(y_2)\,dy_2 + e^{y_1}[d\cos(y_2)]\,dy_2 + e^{y_1}\cos(y_2)\,d^2 y_2 \\ &= e^{y_1}\sin(y_2)\,(dy_1)^2 + e^{y_1}\cos(y_2)\,dy_1\,dy_2 + e^{y_1}\sin(y_2)\,d^2 y_1 \\ &\quad + e^{y_1}\cos(y_2)\,dy_1\,dy_2 - e^{y_1}\sin(y_2)\,(dy_2)^2 + e^{y_1}\cos(y_2)\,d^2 y_2, \end{aligned}$$

where we emphasize that $d^2 y_1$ and $d^2 y_2$ are not 0, because they depend on x_1 and x_2. In fact,

$$dy_1 = x_2^2\,dx_1 + 2x_1x_2\,dx_2, \qquad d^2 y_1 = 4x_2\,dx_1\,dx_2 + 2x_1(dx_2)^2,$$

and

$$dy_2 = 2x_1x_2\,dx_1 + x_1^2\,dx_2, \qquad d^2 y_2 = 2x_2(dx_1)^2 + 4x_1\,dx_1\,dx_2.$$

Inserting these expressions into $d^2 f(y)$ gives the required result. $\|$

11. The Hessian matrix

When we move from vector calculus to matrix calculus, we need an ordering of the functions and of the variables. As motivated in Section 3, 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 already obtained in (33) the extension of the first identification theorem,

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

For the second identification theorem, we obtain

$$d^2 f(X) = (d\operatorname{vec} X)'B(X)\,d\operatorname{vec} X \iff Hf(X) = \frac{B(X) + B(X)'}{2}. \tag{42}$$

Notice that we only provide the Hessian matrix for scalar functions, not for vector or matrix functions, as explained in Section 9.

The commutation matrix, introduced in Section 2.5, has many applications in matrix theory, and it is 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 $\operatorname{tr} A(dX)'BdX$ or $\operatorname{tr} A(dX)BdX$. The following result is then of importance.

Proposition 15. *Let f be a twice differentiable real-valued function of an $n \times q$ matrix X. Then,*

$$d^2 f(X) = \operatorname{tr} A(dX)' B dX \iff Hf(X) = \frac{1}{2}(A' \otimes B + A \otimes B')$$

and

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

Proof. We write

$$\begin{aligned}\operatorname{tr} A(dX)' B dX &= \operatorname{tr}(dX)' B(dX) A = (\operatorname{vec} dX)' \operatorname{vec} B(dX) A \\ &= (\operatorname{vec} dX)'(A' \otimes B) \operatorname{vec} dX = (d \operatorname{vec} X)'(A' \otimes B) d \operatorname{vec} X\end{aligned}$$

and

$$\begin{aligned}\operatorname{tr} A(dX) B dX &= \operatorname{tr}(dX)' B'(dX)' A' = (\operatorname{vec} dX)' \operatorname{vec} B'(dX)' A' \\ &= (\operatorname{vec} dX)'(A \otimes B') \operatorname{vec} dX' = (d \operatorname{vec} X)'(A \otimes B') K_{nq} d \operatorname{vec} X,\end{aligned}$$

using Propositions 5 and 6, and properties of the commutation matrix. The result now follows from Proposition 13, using the fact that

$$(A \otimes B') K_{nq} + K_{qn}(A' \otimes B) = K_{qn}(A' \otimes B + B' \otimes A). \quad \square$$

Exercise 28. Suppose that $d^2 f(X) = 2 \operatorname{tr} A(dX) A\, dX$ and that both A and X are known to be symmetric $n \times n$ matrices. Show that $Hf(X) = 2D_n'(A \otimes A) D_n$.
Solution. This follows from

$$\begin{aligned}\operatorname{tr} A(dX)' A\, dX &= (d \operatorname{vec} X)'(A \otimes A) d \operatorname{vec} X \\ &= (d \operatorname{vech}(X))' D_n'(A \otimes A) D_n\, d \operatorname{vech}(X). \quad \|\end{aligned}$$

12. Four exercises: second derivative

In this section we consider the same functions as in Section 8, but now we obtain the second derivatives.

Exercise 29 (*Exercise* 19 *Cont'd*)**.** Obtain the Hessian of $f(X) = \operatorname{tr} X'AX$.
Solution. We know that $df(X) = \operatorname{tr} C' dX$, where $C = (A + A')X$. The second differential is

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

Hence, the Hessian is $Hf(X) = I_q \otimes (A + A')$.

Exercise 30 (*Exercise* 20 *Cont'd*)**.** Obtain the Hessian of $f(X) = \log |X'X|$, where X has full column rank.
Solution. Since $df(X) = 2 \operatorname{tr} C' dX$, where $C = X(X'X)^{-1}$, the second differential is

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

where $M = I_n - X(X'X)^{-1}X'$. The second equality in this derivation follows from considering $(X'X)^{-1}X'dX$ as a product of three matrices: $(X'X)^{-1}$, X', and dX (a matrix of constants), the third equality uses the differential of the inverse in (32), and the fourth equality separates $dX'X$ into $(dX)'X + X'dX$. Hence, we obtain from Proposition 15,

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

Exercise 31 (*Exercise 21 Cont'd*)**.** Let $f_k(X) = \operatorname{tr} X^k$ $(k = 1, 2, \dots)$. Find the Hessian.
Solution. We have $df_k(X) = k \operatorname{tr} X^{k-1} dX$, and, in particular, $df_1(X) = \operatorname{tr} dX$ and $df_2(X) = 2 \operatorname{tr} X\, dX$. Hence, $d^2 f_1(X) = 0$ and, for $k \geq 2$,

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

This gives $Hf_1(X) = 0$, and, for $k \geq 2$,

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

Exercise 32 (*Exercise 22 Cont'd*)**.** Find the Hessian of the matrix equation $F(X) = AX^{-1}B$, where X is nonsingular.
Solution. We know that $dF(X) = -AX^{-1}(dX)X^{-1}B$. The second differential is

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

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

$$C_{ts} = X^{-1} B e_t e_s' A X^{-1},$$

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

$$d^2 F_{st}(X) = 2 e_s' A X^{-1}(dX)X^{-1}(dX)X^{-1} B e_t = 2 \operatorname{tr} C_{ts}(dX)X^{-1}(dX),$$

and hence

$$HF_{st}(X) = \frac{\partial^2 F_{st}}{(\partial \operatorname{vec} X)(\partial \operatorname{vec} X)'} = K_n (C_{ts}' \otimes X^{-1} + X'^{-1} \otimes C_{ts}).$$

13. Example 2: Maximum likelihood

Consider a sample of $m \times 1$ vectors $y_1, y_2, \dots, y_n$ 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 $(y_1, \dots, y_n)$ 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

$$\Lambda(\mu, \Omega) = -\frac{mn}{2} \log 2\pi - \frac{n}{2} \log |\Omega| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' \Omega^{-1} (y_i - \mu). \tag{43}$$

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

$$\begin{aligned} d\Lambda &= -\frac{n}{2} d \log |\Omega| + \frac{1}{2} \sum_{i=1}^{n} (d\mu)' \Omega^{-1} (y_i - \mu) \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' (d\Omega^{-1}) (y_i - \mu) + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' \Omega^{-1} d\mu \\ &= -\frac{n}{2} d \log |\Omega| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' (d\Omega^{-1}) (y_i - \mu) + \sum_{i=1}^{n} (y_i - \mu)' \Omega^{-1} d\mu \\ &= -\frac{n}{2} \operatorname{tr}(\Omega^{-1} d\Omega + S d\Omega^{-1}) + \sum_{i=1}^{n} (y_i - \mu)' \Omega^{-1} d\mu, \end{aligned} \tag{44}$$

where

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

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

$$\operatorname{tr}\left(\hat{\Omega}^{-1} - \hat{\Omega}^{-1} \hat{S} \hat{\Omega}^{-1}\right) d\Omega = 0, \qquad \sum_{i=1}^{n} (y_i - \hat{\mu})' \hat{\Omega}^{-1} d\mu = 0,$$

for all $d\Omega$ and all $d\mu$. 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 ML estimators are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}, \qquad \hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})'. \tag{45}$$

We note that the condition that Ω is symmetric has not been imposed. But since the solution $\hat{\Omega}$ is symmetric, imposing the condition would have made no difference.

The second differential is obtained by differentiating (44) again. This gives

$$\begin{aligned} d^2\Lambda &= -\frac{n}{2}\operatorname{tr}\left((d\Omega^{-1})d\Omega + (dS)d\Omega^{-1} + Sd^2\Omega^{-1}\right) - n(d\mu)'\Omega^{-1}d\mu \\ &\quad + \sum_{i=1}^{n}(y_i - \mu)'(d\Omega^{-1})d\mu. \end{aligned} \tag{46}$$

From the second differential (46) we can obtain the Hessian matrix by differentiating further, but if one is only interested in the information matrix (minus the expectation of the Hessian matrix), then an important shortcut is possible by taking expectations at this stage. Since $\mathrm{E}(S) = \Omega$ and $\mathrm{E}(dS) = 0$, we obtain

$$\begin{aligned} \mathrm{E}\, d^2\Lambda &= -\frac{n}{2}\operatorname{tr}\left((d\Omega^{-1})d\Omega + \Omega d^2\Omega^{-1}\right) - n(d\mu)'\Omega^{-1}d\mu \\ &= \frac{n}{2}\operatorname{tr}\Omega^{-1}(d\Omega)\Omega^{-1}d\Omega - n\operatorname{tr}(d\Omega)\Omega^{-1}(d\Omega)\Omega^{-1} - n(d\mu)'\Omega^{-1}d\mu \\ &= -\frac{n}{2}\operatorname{tr}\Omega^{-1}(d\Omega)\Omega^{-1}d\Omega - n(d\mu)'\Omega^{-1}d\mu, \end{aligned} \tag{47}$$

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

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

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 (47) without use of the vec operator. Thus, from (47),

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

which implies that the information matrix for μ and $\operatorname{vech}(\Omega)$ takes the form

$$\mathcal{F} = n\begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \frac{1}{2}D_m'(\Omega^{-1}\otimes\Omega^{-1})D_m \end{pmatrix}. \tag{48}$$

The results on the duplication matrix in Section 2.6 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:

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

14. Example 3: Maximum likelihood with parameters in the design matrix

Let us consider the linear model $y = X\beta + u$, where $\mathrm{E}(u) = 0$ and $\operatorname{var}(u) = \Omega(\theta)$. In contrast to the standard linear model we assume, in addition, that X depends on a vector of parameters ψ. We wish to find the ML estimators of β, ψ, and θ, and the information matrix, whose inverse approximates the variance of the ML estimators. This model was considered in Ikefuji et al. (2022).

Under normality, the loglikelihood takes the form

$$\Lambda(\beta,\psi,\theta) = \text{constant} - (1/2)\log|\Omega| - (1/2)(y - X\beta)'\Omega^{-1}(y - X\beta). \tag{49}$$

Maximizing Λ with respect to β and θ is (relatively) easy, while maximization with respect to ψ is more difficult. Upon differentiating $X\beta$ we obtain

$$d(X\beta) = Xd\beta + (dX)\beta = Xd\beta + (\beta'\otimes I_n)Z\,d\psi,$$

where $Z = \partial \operatorname{vec} X / \partial \psi'$. Differentiating the loglikelihood then gives

$$\begin{aligned} d\Lambda &= -(1/2)\operatorname{tr}(\Omega^{-1} d\Omega) + (1/2)(y - X\beta)' \Omega^{-1} (d\Omega) \Omega^{-1} (y - X\beta) \\ &\quad + (y - X\beta)' \Omega^{-1} X d\beta + (y - X\beta)' \Omega^{-1} (dX)\beta, \end{aligned} \tag{50}$$

from which we obtain the first-order conditions

$$\begin{aligned} (y - X\beta)' \Omega^{-1} X d\beta &= 0, \\ (y - X\beta)' \Omega^{-1} (d\Omega) \Omega^{-1} (y - X\beta) &= \operatorname{tr}(\Omega^{-1} d\Omega), \\ (y - X\beta)' \Omega^{-1} (dX)\beta &= 0, \end{aligned}$$

for β, θ, and ψ, respectively. This implies that $\hat{\beta}$ takes the simple form

$$\hat{\beta}(\psi, \theta) = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y, \tag{51}$$

so we can concentrate the likelihood with respect to β. The concentrated loglikelihood is

$$\Lambda_c(\psi, \theta) = \text{constant} - (1/2)\log|\Omega| - (1/2)\hat{u}' \Omega^{-1} \hat{u},$$

where $\hat{u} = y - X\hat{\beta} = y - X(X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y$. We obtain $\hat{\psi}$ and $\hat{\theta}$ by maximizing Λ_c, and then $\hat{\beta}$ through (51).

One way to maximize Λ_c is to perform a grid search over ψ. Fixing $\psi = \psi_0$, we have $X = X(\psi_0)$ and $d\psi = 0$. Then,

$$\begin{aligned} d\hat{\beta} &= [d(X' \Omega^{-1} X)^{-1}] X' \Omega^{-1} y + (X' \Omega^{-1} X)^{-1} d(X' \Omega^{-1} y) \\ &= -(X' \Omega^{-1} X)^{-1} X' \Omega^{-1} (d\Omega) \Omega^{-1} \hat{u}, \end{aligned}$$

and hence

$$\begin{aligned} d\Lambda_c(\psi_0, \theta) &= -(1/2)\operatorname{tr}(\Omega^{-1} d\Omega) + (1/2)\hat{u}' \Omega^{-1} (d\Omega) \Omega^{-1} \hat{u} \\ &\quad - \hat{u}' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} (d\Omega) \Omega^{-1} \hat{u}. \end{aligned}$$

Solving θ from $d\Lambda_c(\psi_0, \theta) = 0$ gives $\hat{\theta}(\psi_0)$ and hence $\Lambda_c(\psi_0, \hat{\theta}(\psi_0))$. Performing a grid search over ψ we find $\hat{\psi}$ where $\Lambda_c(\psi, \hat{\theta}(\psi))$ is maximized. Given $\hat{\psi}$ and $\hat{\theta}$ we then compute $\hat{\beta}$ from (51).

To obtain the information matrix we take the differential of $d\Lambda$ in (50). This gives, letting $\mu = X\beta$,

$$\begin{aligned} d^2\Lambda &= (1/2)\operatorname{tr}(\Omega^{-1} d\Omega)^2 - (y - X\beta)' \Omega^{-1} (d\Omega) \Omega^{-1} (d\Omega) \Omega^{-1} (y - X\beta) \\ &\quad - (d\mu)' \Omega^{-1} (d\mu) - 2(y - X\beta)' \Omega^{-1} (d\Omega) \Omega^{-1} (d\mu) + (y - X\beta)' \Omega^{-1} (d^2\mu) \\ &\quad - (1/2)\operatorname{tr}(\Omega^{-1} d^2\Omega) + (1/2)(y - X\beta)' \Omega^{-1} (d^2\Omega) \Omega^{-1} (y - X\beta). \end{aligned} \tag{52}$$

Minus the expectation of the second differential takes the simple form

$$-\mathrm{E}(d^2\Lambda) = (1/2)\operatorname{tr}(\Omega^{-1} d\Omega)^2 + (d\mu)' \Omega^{-1} (d\mu),$$

which implies that the information matrix is block-diagonal in (β, ψ) and θ. Therefore we do not have to take the variance of the ML estimator $\hat{\theta}$ into account when calculating the variance of the ML estimators $(\hat{\beta}, \hat{\psi})$. Now writing

$$\operatorname{tr}(\Omega^{-1} d^2\Omega) = (d \operatorname{vec} \Omega)' (\Omega^{-1} \otimes \Omega^{-1})(d \operatorname{vec} \Omega) = (d\theta)' \mathcal{F}_{\theta\theta}\, d\theta,$$

with

$$\mathcal{F}_{\theta\theta} = \left(\frac{\partial \operatorname{vec} \Omega}{\partial \theta'}\right)' (\Omega^{-1} \otimes \Omega^{-1}) \left(\frac{\partial \operatorname{vec} \Omega}{\partial \theta'}\right),$$

and

$$(d\mu)' \Omega^{-1} (d\mu) = \begin{pmatrix} d\beta \\ d\psi \end{pmatrix}' \begin{pmatrix} \mathcal{F}_{\beta\beta} & \mathcal{F}_{\beta\psi} \\ \mathcal{F}_{\psi\beta} & \mathcal{F}_{\psi\psi} \end{pmatrix} \begin{pmatrix} d\beta \\ d\psi \end{pmatrix},$$

with

$$\mathcal{F}_{\beta\beta} = X' \Omega^{-1} X, \qquad \mathcal{F}_{\beta\psi} = \mathcal{F}'_{\psi\beta} = X' \left(\beta' \otimes \Omega^{-1}\right) Z,$$

and

$$\mathcal{F}_{\psi\psi} = Z' \left(\beta\beta' \otimes \Omega^{-1}\right) Z,$$

we obtain the information matrix

$$\mathcal{F} = \begin{pmatrix} \mathcal{F}_{\beta\beta} & \mathcal{F}_{\beta\psi} & 0 \\ \mathcal{F}_{\psi\beta} & \mathcal{F}_{\psi\psi} & 0 \\ 0 & 0 & \mathcal{F}_{\theta\theta} \end{pmatrix},$$

whose inverse $\mathcal{F}^{-1}$ approximates the variance matrix of the ML estimators.

15. Example 4: The Eckart–Young theorem

My next example is somewhat related to restricted least squares (Section 6), but instead of approximating a vector Ab, we wish to approximate A itself. More specifically, I wish to approximate a given $m \times n$ matrix A by a matrix XZ', such that

$$f(X,Z) = \operatorname{tr}(A - XZ')(A - XZ')' \tag{53}$$

is minimized. Let r be a given (small) integer such that X has dimension $m \times r$ and Z has dimension $n \times r$, where the latter is normalized by the semi-orthogonality condition $Z'Z = I_r$.

The Eckart–Young theorem (Eckart and Young, 1936, Magnus and Neudecker, 2019, Theorem 17.7) tells us that the minimum of f is obtained when Z contains the eigenvectors associated with the r *largest* eigenvalues of $A'A$ and $X = AZ$. The 'best' approximation $\tilde{A}$ (of rank r) to A is then $\tilde{A} = AZZ'$, and the constrained minimum of f is the sum of the $n - r$ *smallest* eigenvalues of $A'A$.

To prove the Eckart–Young theorem, define the Lagrangian function

$$\mathcal{L}(X,Z) = \frac{1}{2}\operatorname{tr}(A - XZ')(A - XZ')' - \frac{1}{2}\operatorname{tr} L(Z'Z - I_r), \tag{54}$$

where L is a symmetric $r \times r$ matrix of Lagrange multipliers (see Exercise 18). Differentiating $\mathcal{L}$, we obtain

$$\begin{aligned} d\mathcal{L}(X,Z) &= \operatorname{tr}(A - XZ')d(A - XZ')' - \frac{1}{2}\operatorname{tr} L\left((dZ)'Z + Z'dZ\right) \\ &= -\operatorname{tr}(A - XZ')Z(dX)' - \operatorname{tr}(A - XZ')(dZ)X' - \operatorname{tr} LZ'dZ \\ &= -\operatorname{tr}(A - XZ')Z(dX)' - \operatorname{tr}(X'A - X'XZ' + LZ')dZ. \end{aligned}$$

The first-order conditions are therefore

$$\begin{aligned} (A - XZ')Z &= 0, \\ X'A - X'XZ' + LZ' &= 0, \\ Z'Z &= I_r. \end{aligned} \tag{55}$$

These conditions give $X = AZ$ and hence $L = X'X - X'AZ = 0$, so that we arrive at the equation

$$(A'A)Z = Z(Z'A'AZ). \tag{56}$$

Now, let P be an orthogonal $r \times r$ matrix such that

$$P'(Z'A'AZ)P = \Lambda_1,$$

where Λ_1 is a diagonal $r \times r$ matrix containing the eigenvalues of $Z'A'AZ$ on its diagonal. Let $T_1 = ZP$. Then (56) can be written as $A'AT_1 = T_1\Lambda_1$, where T_1 is a semi-orthogonal $n \times r$ matrix (that is, it satisfies $T_1'T_1 = I_r$) that diagonalizes $A'A$, and the r diagonal elements in Λ_1 are eigenvalues of $A'A$. (The matrix $A'A$ has n eigenvalues and r of these are contained in Λ_1.)

Given $X = AZ$, we have

$$(A - XZ')(A - XZ')' = A(I - ZZ')A',$$

and thus

$$\operatorname{tr}(A - XZ')(A - XZ')' = \operatorname{tr} A'A - \operatorname{tr}\Lambda_1. \tag{57}$$

To minimize (57), we must maximize $\operatorname{tr}\Lambda_1$. Hence Λ_1 must contain the r largest eigenvalues of $A'A$, and T_1 contains eigenvectors associated with these r eigenvalues. The 'best' approximation to A is then

$$XZ' = AZZ' = AT_1T_1',$$

so that an optimal choice is $Z = T_1$, $X = AT_1$. From (57), it is clear that the value of the constrained minimum is the sum of the $n - r$ smallest eigenvalues of $A'A$.

16. Tacit knowledge

When you want to become a carpenter or a doctor or a chef, you follow courses, read books, and pass exams, and at some point you become a trainee in a studio, clinic, or restaurant. There you learn important things that are not available in books and were not taught to you in courses. Such knowledge is called 'tacit knowledge'. Let me try, in this final section, to write down a few things that I have learned over the years in the hope that they may be of use to the reader. The following suggestions refer to matrix algebra in general.

- Think of matrices as units of a higher-order algebra. Do not think in terms of the elements of a matrix.
- When proving a theorem about $n \times n$ matrices, try to prove it first for $n = 2$ and $n = 3$. If it works for $n = 2$, then this is good news, but it is no guarantee. But if it works for $n = 3$, then it probably works in general.

- When a matrix is symmetric, set it equal to a diagonal matrix, and see if the theorem works in that case.
- If you need to prove that $A = B$, it is almost always easier to prove that $C = A - B = 0$. There are many ways to prove that $C = 0$, but the method of showing that $c_{ij} = 0$ for all i and j is probably not the most efficient. Maybe you can prove that $\operatorname{tr} C'C = 0$, which is equivalent.
- If we have a matrix A and a matrix B which is the same as A except that (for example) its last column is missing, then do not write $B = \tilde{A}$, but write B explicitly in terms of A, in the present case $B = AE$, where E is a matrix of zeros and ones.

The next suggestions refer specifically to matrix calculus.

- Always use the correct definition of matrix derivative. The derivative of $\operatorname{tr} X$ is not the identity matrix but $(\operatorname{vec} I)'$, a row vector.
- When you need the Hessian, do not start with the first derivative, but with the first differential. Then obtain the second differential, then the Hessian (see the exercises in Section 12).
- Remember that Cauchy invariance does not hold for second differentials. I usually avoid the chain rule for second differentials, and write the first differential in terms of the eventual matrices, unless there is linearity (symmetry, diagonality), which can be safely added at the end.
- If $d^2 f(x) = (dx)' B(x)\, dx$, remember to 'symmetrize' the matrix B, because B will not be symmetric in general. The Hessian is then $H f(x) = (B(x) + B(x)')/2$.

Declaration of competing interest

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I thank Serena Ng for alerting me to the 'how-to' paper series, the editors for supporting the idea of a tutorial on matrix calculus, and two unusually constructive referees and Yannick van Etten for valuable comments. As always, Henk Pijls has been my unfailing 'helpdesk' for mathematical questions of all types. My publisher John Wiley gracefully permitted me to freely use material from *Matrix Differential Calculus*, 3rd edition.

References

Note: The references below are in no way representative of the literature on matrix calculus, as I have only included papers and books that I refer to in the current paper. The third edition of *Matrix Differential Calculus* provides an up-to-date bibliography.

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218.
Ikefuji, M., Laeven, R.J.A., Magnus, J.R., Yue, Y., 2022. Earthquake risk embedded in property prices: Evidence from five Japanese cities. J. Amer. Statist. Assoc. 117, 82–93.
Magnus, J.R., 1988. Linear Structures. Oxford University Press, New York.
Magnus, J.R., 2010. On the concept of matrix derivative. J. Multivariate Anal. 101, 2200–2206.
Magnus, J.R., 2024. Matrix derivatives: Why and where did it go wrong? IMAGE. Bull. Int. Linear Algebra Soc. 72 (Spring), 3–8.
Magnus, J.R., Neudecker, H., 2019. Matrix Differential Calculus with Applications in Statistics and Econometrics, third ed. John Wiley, Chichester/New York.