FACULTY OF ACTUARIAL SCIENCE & ECONOMETRICS

A & E REPORT

REPORT AE 19/84

MATRIX DIFFERENTIAL CALCULUS WITH APPLICATIONS TO SIMPLE, HADAMARD, AND KRONECKER PRODUCTS

J.R. Magnus
H. Neudecker

University of Amsterdam

Title: Matrix differential calculus with applications to simple, Hadamard, and Kronecker products

Authors : J.R. Magnus
(London School of Economics)
H. Neudecker
(University of Amsterdam)

Address: University of Amsterdam
Faculty of Actuarial Science & Econometrics
Jodenbreestraat 23
1011 NH Amsterdam

Date: September 1984

Series and number: AE 19/84

Pages: 28

Price: No charge

JEL subject classification: 213

Keywords: Matrix differential calculus; Kronecker product; Commutation matrix; Hadamard product.

Abstract: A method of matrix differential calculus based on differentials is presented. Several definitions of matrix derivatives are examined, but only one is retained. The resulting formulae are applied to matrix functions of a more complicated character.

1. Introduction

In mathematical psychology (among other disciplines) matrix calculus has become an important tool. Several definitions are in use for the derivative of a matrix function F(X) with respect to its matrix argument X. (See Dwyer and MacPhail (1948), Dwyer (1967), Neudecker (1969), McDonald and Swaminathan (1973), MacRae (1974), Balestra (1976), Nel (1980), Rogers (1980), and Polasek (1984). We shall see that only one of these definitions is appropriate. Also, several prodecures exist for a calculus of functions of matrices. We shall argue that the procedure based on differentials is superior over other methods of differentiation.

These, then are the two purposes of this paper. We have been inspired to write this paper by Bentler and Lee's (1978) note. Although they use the right type of matrix derivative (which is similar but not the same as our definition), Bentler and Lee present a very confused chain rule based on "mathematically independent variables", and their procedure to obtain Jacobian matrices for matrix functions is unsatisfactory. (We note in passing their unusual definition of a differentiable function, as one whose partial derivatives exist.) In the present paper we give a completely satisfactory chain rule for matrix functions, and show that the approach via differentials is elegant, short, and easy.

Keeping within the limited space of a journal article we restrict ourselves to first-order differentials and thus to first derivatives. The presentation will culminate in the so-called First Identification Theorem. This will be the basic theorem from which the results on derivatives will follow.

The following notation is used. An $m \times n$ matrix is one having m rows and n columns; A' denotes the transpose of A, $\operatorname{tr} A$ denotes its trace (if A is square), and A^{-1} its inverse (if A is nonsingular). $\mathbb{R}^{m \times n}$ is the class of real $m \times n$ matrices and $\mathbb{R}^n$ the class of real $n \times 1$ vectors, so that $\mathbb{R}^n \equiv \mathbb{R}^{n \times 1}$. The $n \times n$ identity matrix is denoted I_n. Let A be an $m \times n$ matrix. Then $\operatorname{vec} A$ is the $mn \times 1$ vector that stacks the columns of A one underneath the other, and $\hat{A}$ is the $mn \times mn$ diagonal matrix displaying the elements of $\operatorname{vec} A$ along its diagonal. The commutation matrix K_{mn} (Magnus and Neudecker (1979)) is the $mn \times mn$ matrix which transforms $\operatorname{vec} A$ into $\operatorname{vec} A'$:

$$K_{mn} \operatorname{vec} A = \operatorname{vec} A'. \tag{1}$$

$A * B$ denotes the Hadamard product $(a_{ij} b_{ij})$, and $A \otimes B$ denotes the Kronecker product $(a_{ij} B)$. It is easy to see that

$$\operatorname{vec} ab' = b \otimes a \tag{2}$$

for any two vectors a and b. We shall use (2) in section 10.

2. Bad notation

Let us start with some remarks on notation. If F is a differentiable m×p matrix function of an n×q matrix of variables X, then the question arises how to display the mnpq partial derivatives of F with respect to X. Obviously, this can be done in many ways. One possible ordering of the partial derivatives, very popular but unsuitable for theoretical work as we shall see shortly, is contained in the following two definitions.

Definition 1. Let ϕ be a differentiable real-valued function of an n×q matrix of real variables $X = (x_{ij})$. Then the symbol $\partial\phi(X)/\partial X$ denotes the n×q matrix

$$\frac{\partial\phi(X)}{\partial X} = \begin{pmatrix} \partial\phi/\partial x_{11} & \cdots & \partial\phi/\partial x_{1q} \\ \vdots & & \vdots \\ \partial\phi/\partial x_{n1} & \cdots & \partial\phi/\partial x_{nq} \end{pmatrix}. \tag{1}$$

Definition 2. Let $F = (f_{st})$ be a differentiable m×p real matrix function of an n×q matrix of real variables X. Then the symbol $\partial F(X)/\partial X$ denotes the mn×pq matrix

$$\frac{\partial F(X)}{\partial X} = \begin{pmatrix} \partial f_{11}/\partial X & \cdots & \partial f_{1p}/\partial X \\ \vdots & & \vdots \\ \partial f_{m1}/\partial X & \cdots & \partial f_{mp}/\partial X \end{pmatrix}. \tag{2}$$

Before we criticise Definition 2, let us list some of its good (notational) points. Two very pleasant properties are: (i) if F is a matrix function of just one variable ξ, then $\partial F(\xi)/\partial\xi$ has the same order as $F(\xi)$; and (ii) if ϕ is a scalar function of a matrix of variables X, then $\partial\phi(X)/\partial X$ has the same order as X. In particular, if ϕ is a scalar function of a column-vector x, then $\partial\phi/\partial x$ is a column-vector and $\partial\phi/\partial x'$ a row-vector. Another consequence of the definition is that it allows us to display the mn partial derivatives of an $m\times 1$ vector function f(x), where x is an $n\times 1$ vector of variables, in four ways, namely as $\partial f/\partial x'$ (an $m\times n$ matrix), $\partial f'/\partial x$ (an $n\times m$ matrix), $\partial f/\partial x$ (an $mn\times 1$ vector), or as $\partial f'/\partial x'$ (an $1\times mn$ vector).

These are undoubtedly strong notational virtues of Definition 2. To see what is wrong with the definition from a theoretical point of view (see also Pollock (1984) on this point) let us consider the identity function F(X) = X, where X is an $n\times q$ matrix of real variables. We obtain from Definition 2

$$\frac{\partial F(X)}{\partial X} = (\text{vec } I_n)(\text{vec } I_q)', \tag{3}$$

a matrix of rank one. The Jacobian matrix of the identity function is, of course, I_{nq}, the $nq\times nq$ identity matrix. Hence Definition 2 does not give us the Jacobian matrix of the function F, and, indeed, the rank of the Jacobian matrix is not given by the rank of $\partial F(X)/\partial X$. This implies -- and this cannot be stressed enough -- that the matrix (2) displays the partial derivatives, but nothing more. In particular, the determinant of $\partial F(X)/\partial X$ has no interpretation, and (very important for practical work also) a useful chain rule does not exist.

There exists another definition, equally unsuitable, which is based not on $\partial\phi(X)/\partial X$, but on

$$\frac{\partial F(X)}{\partial x_{ij}} = \begin{pmatrix} \partial f_{11}(X)/\partial x_{ij} & \cdots & \partial f_{1p}(X)/\partial x_{ij} \\ \vdots & & \vdots \\ \partial f_{m1}(X)/\partial x_{ij} & \cdots & \partial f_{mp}(X)/\partial x_{ij} \end{pmatrix}. \tag{4}$$

Definition 3. Let F be a differentiable m×p matrix function of an n×q matrix of real variables $X = (x_{ij})$. Then the symbol $\partial F(X)//\partial X$ denotes the mn×pq matrix

$$\frac{\partial F(X)}{\partial X} = \begin{pmatrix} \partial F(X)/\partial x_{11} & \cdots & \partial F(X)/\partial x_{1q} \\ \vdots & & \vdots \\ \partial F(X)/\partial x_{n1} & \cdots & \partial F(X)/\partial x_{nq} \end{pmatrix} . \qquad (5)$$

From a theoretical viewpoint Definitions 2 and 3 are equally bad. Definition 3 has, however, one practical advantage over Definition 2 in that the expressions $\partial F(X)/\partial x_{ij}$ are much easier to evaluate than $\partial f_{st}(X)/\partial X$.

After these critical remarks, let us turn quickly to the only natural and viable generalization of the notion of a Jacobian matrix of a vector function to a Jacobian matrix of a matrix function.

3. Good notation

We recommend to proceed in the following way. Let ϕ be a scalar function of an $n\times 1$ vector x. Then the $1\times n$ vector

$$\nabla\phi(x) = \frac{\partial\phi(x)}{\partial x'} \tag{1}$$

is called the gradient of ϕ. If f is an $m\times 1$ vector function of x, then the $m\times n$ matrix

$$\nabla f(x) = \begin{pmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_m(x) \end{pmatrix} = \frac{\partial f(x)}{\partial x'} \tag{2}$$

is called the Jacobian matrix of f. Since (1) is just a special case of (2), the double use of the ∇-symbol is permitted. Generalizing these concepts to matrix functions of matrices, we arrive at

Definition 4. Let F be a differentiable $m\times p$ real matrix function of an $n\times q$ matrix of real variables X. The Jacobian matrix of F at X (the gradient vector, if $m=p=1$) is the $mp\times nq$ matrix

$$\nabla F(X) = \frac{\partial \operatorname{vec} F(X)}{\partial (\operatorname{vec} X)'} \;. \;\| \tag{3}$$

Thus ∇F, ∇f, and $\nabla\phi$ are all defined. We refer to ∇F and ∇f as Jacobian matrices, and to $\nabla\phi$ as a gradient vector.

It is important to notice that $\nabla F(X)$ and $\partial F(X)/\partial X$ contain the same $mnpq$ partial derivatives, but in a different pattern. Indeed, the orders of the two matrices are different ($\nabla F(X)$ is of the order $mp\times nq$, while $\partial F(X)/\partial X$ is of the order $mn\times pq$), and, more important, their ranks are in general different.

Since $\nabla F(X)$ is a straightforward matrix generalization of the traditional definition of the Jacobian matrix $\partial f(x)/\partial x'$, all properties of Jacobian matrices are preserved. In particular, questions relating to functions with non-zero Jacobian determinant at certain points remain meaningful, as does the chain rule.

Definition 4 reduces the study of matrix functions of matrices to the study of vector functions of vectors, since it allows F(X) and X only in their vectorized forms vec F and vec X. As a result, the unattractive expressions

$$\frac{\partial F(X)}{\partial X}, \quad \frac{\partial F(x)}{\partial x}, \quad \text{and} \quad \frac{\partial f(X)}{\partial X} \tag{5}$$

are not needed. The same holds, in principle, for the expressions

$$\frac{\partial \phi(X)}{\partial X} \quad \text{and} \quad \frac{\partial F(\xi)}{\partial \xi}, \tag{6}$$

since these can be replaced by

$$\nabla\phi(X) = \frac{\partial \phi(X)}{\partial (\text{vec } X)'} \quad \text{and} \quad \nabla F(\xi) = \frac{\partial \, \text{vec } F(\xi)}{\partial \xi}. \tag{7}$$

However, the idea of arranging the partial derivatives of $\phi(X)$ and $F(\xi)$ into a matrix (rather than a vector) is rather appealing and sometimes useful, so we retain the expressions in (6).

4. The differential

Fundamental to our approach is the concept of a differential. In the one-dimensional case, the equation

$$\lim_{u \to o} \frac{\phi(c+u) - \phi(c)}{u} = \phi'(c) \qquad (1)$$

defining the derivative at c is equivalent to the equation

$$\phi(c+u) = \phi(c) + u\phi'(c) + r_c(u), \qquad (2)$$

where the remainder $r_c(u)$ is of smaller order than u as $u \to o$, that is,

$$\lim_{u \to o} \frac{r_c(u)}{u} = 0\,. \qquad (3)$$

Equation (2) is called the first-order Taylor formula. If for the moment we think of the point c as fixed and the increment u as variable, then the increment of the function, that is the quantity $\phi(c+u) - \phi(c)$, consists of two terms, namely a part $u\phi'(c)$ which is proportional to u and an "error" which can be made as small as we please relative to u by making u itself small enough. Thus the smaller the interval about the point c which we consider, the more accurately is the function $\phi(c+u)$ -- which is a function of u -- represented by its affine linear part $\phi(c) + u\phi'(c)$. We now define

$$d\phi(c;u) = u\phi'(c) \qquad (4)$$

as the (first) differential of ϕ at c with increment u.

The notation $d\phi(c;u)$ rather than $d\phi(c,u)$ emphasizes the different roles of c and u. The first point, c, must be a point where $\phi'(c)$ exists, whereas the second point, u, is an arbitrary point in $\mathbb{R}$.

Although the concept of differential is as a rule only used when u is small, there is in principle no need to restrict u in any way. In particular, the differential $d\phi(c;u)$ is a number which has nothing to do with infinitely small quantities.

Conversely, if there exists a quantity α, depending on c but not on u, such that

$$\phi(c+u) = \phi(c) + u\alpha + r(u) , \tag{5}$$

where $r(u)/u$ tends to o with u, that is if we can approximate $\phi(c+u)$ by an affine linear function (in u) such that the difference between the function and the approximation function vanishes to a higher order than the increment u, then ϕ is differentiable at c. The quantity α must then be taken as the derivative $\phi'(c)$. We see this immediately if we rewrite condition (5) in the form

$$\frac{\phi(c+u) - \phi(c)}{u} = \alpha + \frac{r(u)}{u} \tag{6}$$

and then let u tend to o. Differentiability of a function with respect to a variable and the possibility of approximating a function by means of an affine linear function in this way are therefore equivalent properties.

These ideas can be extended in a perfectly natural way to vector functions of two or more variables.

Definition 5. Let $f : S \to \mathbb{R}^m$ be a function defined on a set S in $\mathbb{R}^n$. Let c be an interior point of S, and let $B(c;r)$ be an n-ball lying in S. Let u be a point in $\mathbb{R}^n$ with $\|u\| < r$, so that $c + u \in B(c;r)$. If there exists a real $m \times n$ matrix A, depending on c but not on u, such that

$$f(c+u) = f(c) + A(c)u + r_c(u) \qquad (7)$$

for all $u \in \mathbb{R}^n$ with $\|u\| < r$, and

$$\lim_{u \to 0} \frac{r_c(u)}{\|u\|} = 0 \, , \qquad (8)$$

then the function f is said to be <u>differentiable at c</u>; the $m \times 1$ vector

$$df(c;u) = A(c)u \qquad (9)$$

is then called the <u>(first) differential of f at c with increment u</u>. If f is differentiable at every point of an open subset E of S, we say f is <u>differentiable on E.</u> ||

In other words, f is differentiable at the point c if f(c+u) can be approximated by an affine linear function of u. Note that a function f can only be differentiated at an interior point or on an open set.

5. Uniqueness and Existence of the differential

The uniqueness of the differential (if it exists) is contained in the following theorem which is easily established from Definition 5.

Theorem 1 (uniqueness). Let $f : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$, be differentiable at a point $c \in S$ with differential $df(c;u) = A(c)u$. Suppose a second matrix $A^*(c)$ exists such that $df(c;u) = A^*(c)u$. Then $A(c) = A^*(c)$. ||

Now consider the limit

$$\lim_{t \to 0} \frac{f_i(c+te_j) - f_i(c)}{t} . \tag{1}$$

When this limit exists, it is called the partial derivative of f_i with respect to the j^{th} coordinate (or the j^{th} partial derivative of f_i) and is denoted by $D_j f_i(c)$. If f is differentiable at c, then all partial derivatives $D_j f_i(c)$ exist. But the existence of all partial derivatives at c does not imply the existence of the differential at c. Indeed, neither the continuity of all partial derivatives at a point c nor the existence of all partial derivatives in some n-ball B(c) is, in general, a sufficient condition for differentiability. With this knowledge the reader can now appreciate the following theorem.

Theorem 2 (existence). Let $f : S \to \mathbb{R}^m$ be a function defined on a set S in $\mathbb{R}^n$, and let c be an interior point of S. If each of the partial derivatives $D_j f_i$ exists in some n-ball B(c) and is continuous at c, then f is differentiable at c (and hence the differential $df(c;u)$ exists).

6. The First Identification Theorem

If f is differentiable at c, then a matrix A(c) exists such that for all $\|u\| < r$,

$$f(c+u) = f(c) + A(c)u + r_c(u) , \qquad (1)$$

where $r_c(u)/\|u\| \to 0$ as $u \to 0$. It turns out that the elements $a_{ij}(c)$ of the matrix A(c) are, in fact, precisely the partial derivatives $D_j f_i(c)$. This, in conjunction with the uniqueness theorem (Theorem 1) establishes the following central result.

Theorem 3 (First Identification Theorem). Let $f : S \to \mathbb{R}^m$ be a vector function defined on a set S in $\mathbb{R}^n$, and differentiable at an interior point c of S. Let u be a real n×1 vector. Then

$$df(c;u) = (\nabla f(c))u , \qquad (2)$$

where $\nabla f(c)$ is an m×n matrix whose elements $D_j f_i(c)$ are the partial derivatives of f evaluated at c (the Jacobian matrix). Conversely, if A(c) is a matrix such that

$$df(c;u) = A(c)u \qquad (3)$$

for all real n×1 vectors u, then $A(c) = \nabla f(c)$. ||

The great practical value of Theorem 3 lies in the fact that if f is differentiable at c and we have found a differential df at c, then the Jacobian matrix $\nabla f(c)$ can be immediately determined.

Some caution is required when interpreting equation (2). Although the right side of (2) exists if all the partial derivatives $D_j f_i(c)$ exist, the left side of (2) exists only if the differential of f at c exists, i.e., if f is differentiable at c.

7. The chain rule and Cauchy-invariance

An important result of multivariable calculus, which we present without proof is

Theorem 4 (Chain Rule). Let S be a subset of $\mathbb{R}^n$, and assume that $f : S \to \mathbb{R}^m$ is differentiable at an interior point c of S. Let T be a subset of $\mathbb{R}^m$ such that $f(x) \in T$ for all $x \in S$, and assume that $g : T \to \mathbb{R}^p$ is differentiable at an interior point $b = f(c)$ of T. Then the composite function $h : S \to \mathbb{R}^p$ defined by

$$h(x) = g(f(x)) \tag{1}$$

is differentiable at c, and

$$\nabla h(c) = \left(\nabla g(b)\right)\left(\nabla f(c)\right) . \quad \| \tag{2}$$

The chain rule relates the partial derivatives of a composite function $h = g \circ f$ to the partial derivatives of g and f. We shall now discuss an immediate consequence of the chain rule, which relates the differential of h to the differentials of g and f. This result (known as Cauchy's rule of invariance is particularly useful in performing computations with differentials.

Let $h = g \circ f$ be a composite function, as before, such that

$$h(x) = g(f(x)), \qquad x \in S . \tag{3}$$

If f is differentiable at c and g is differentiable at $b = f(c)$, then h is differentiable at c with

$$dh(c;u) = \left(\nabla h(c)\right)u . \tag{4}$$

Using the chain rule, (4) becomes

$$dh(c;u) = (\nabla g(b))(\nabla f(c))u$$

$$= (\nabla g(b))df(c;u) = dg(b;df(c;u)). \qquad (5)$$

We have thus proved

Theorem 5 (Cauchy's Rule of Invariance). If f is differentiable at c and g is differentiable at b = f(c), then the differential of the composite function h = g o f is

$$dh(c;u) = dg(b;df(c;u)) \qquad (6)$$

for every u in $\mathbb{R}^n$. ||

Cauchy's rule of invariance justifies the use of a simpler notation for differentials in practical applications, which adds greatly to the ease and elegance of performing computations with differentials. We shall explain and apply this simpler notation in sections 9 and 10.

8. Matrix functions

To extend the calculus of vector functions to matrix functions is straightforward. Let us consider a matrix function $F : S \to \mathbb{R}^{m\times p}$ defined on a set S in $\mathbb{R}^{n\times q}$. That is, F maps an $n\times q$ matrix X into an $m\times p$ matrix $F(X)$.

Definition 6. Let $F : S \to \mathbb{R}^{m\times p}$ be a matrix function defined on a set S in $\mathbb{R}^{n\times q}$. Let C be an interior point of S, and let $B(C;r) \subset S$ be a ball with centre C and radius r. Let U be a point in $\mathbb{R}^{n\times q}$ with $\|U\| < r$, so that $C + U \in B(C;r)$. If there exists a real $mp\times nq$ matrix A, depending on C but not on U, such that

$$\text{vec}\, F(C+U) = \text{vec}\, F(C) + A(C)\, \text{vec}\, U + \text{vec}\, R_C(U) \qquad (1)$$

for all $U \in \mathbb{R}^{n\times q}$ with $\|U\| < r$, and

$$\lim_{U\to 0} \frac{R_C(U)}{\|U\|} = 0\ , \qquad (2)$$

then the function F is said to be <u>differentiable at C</u>; the $m\times p$ matrix $dF(C;U)$ defined by

$$\text{vec}\, dF(C;U) = A(C)\, \text{vec}\, U \qquad (3)$$

is then called the <u>(first) differential of F at C with increment U</u>.

Note. Recall that the <u>norm</u> of a real matrix X is defined by

$$\|X\| = (\text{tr}\, X'X)^{\frac{1}{2}} \qquad (4)$$

and a <u>ball</u> in $\mathbb{R}^{n\times q}$ by

$$B(C;r) = \{X : X \in \mathbb{R}^{n\times q},\ \|X-C\| < r\}\ . \qquad (5)$$

In view of Definition 6 all calculus properties of matrix functions follow immediately from the corresponding properties of vector functions, because, instead of the matrix function F, we can consider the vector function $f : \text{vec } S \to \mathbb{R}^{mp}$ defined by

$$f(\text{vec } X) \; = \; \text{vec } F(X). \tag{6}$$

It is easy to see that the differentials of F and f are related by

$$\text{vec } dF(C;U) \; = \; df(\text{vec } C; \text{vec } U). \tag{7}$$

This justifies Definition 4, according to which the Jacobian matrix of F at C is

$$\nabla F(C) \; = \; \nabla f(\text{vec } C). \tag{8}$$

This is an $mp \times nq$ matrix, whose ij^{th} element is the partial derivative of the i^{th} component of vec F(X) with respect to the j^{th} element of vec X, evaluated at X = C.

The following three theorems are now straightforward generalizations of Theorems 3-5.

Theorem 6 (First Identification Theorem for matrix functions). Let $F : S \to \mathbb{R}^{m\times p}$ be a matrix function defined on a set S in $\mathbb{R}^{n\times q}$, and differentiable at an interior point C of S. Then

$$\text{vec } dF(C;U) \; = \; A(C) \text{ vec } U \tag{9}$$

for all $U \in \mathbb{R}^{n\times q}$, if and only if

$$\nabla F(C) \; = \; A(C) \; . \tag{10}$$

Theorem 7 (Chain Rule for matrix functions). Let S be a subset of $\mathbb{R}^{n\times q}$, and assume that $F : S \to \mathbb{R}^{m\times p}$ is differentiable at an interior point C of S. Let T be a subset of $\mathbb{R}^{m\times p}$ such that $F(X) \in T$ for all $X \in S$, and assume that $G : T \to \mathbb{R}^{r\times s}$ is differentiable at an interior point $B = F(C)$ of T. Then the composite function $H : S \to \mathbb{R}^{r\times s}$ defined by

$$H(X) = G\big(F(X)\big) \tag{11}$$

is differentiable at C, and

$$\nabla H(C) = \big(\nabla G(B)\big)\big(\nabla F(C)\big) \; . \tag{12}$$

<u>Theorem 8 (Cauchy's Rule of Invariance for matrix functions)</u>. If F is differentiable at C and G is differentiable at $B = F(C)$, then the differential of the composite function $H = G\circ F$ is

$$dH(C;U) = dG\big(B;dF(C;U)\big) \tag{13}$$

for every U in $\mathbb{R}^{n\times q}$.

9. Fundamental rules

We shall now discuss rules for finding differentials, so that we can apply the First Identification Theorem in its two variants. The rules for real-valued (scalar) functions are well-known. We have for real-valued functions u and v, and a real constant α:

$$d\alpha = 0 \tag{1}$$

$$d(\alpha u) = \alpha du \tag{2}$$

$$d(u+v) = du + dv \tag{3}$$

$$d(uv) = (du)v + udv \tag{4}$$

$$d(1/u) = -(1/u^2)du \qquad (u \neq 0) \tag{5}$$

$$du^\alpha = \alpha u^{\alpha-1}du \quad \text{(under obvious restrictions on } \alpha \text{ and } u) \tag{6}$$

$$d\log u = u^{-1}du \qquad (u > 0) \tag{7}$$

$$de^u = e^u du \tag{8}$$

$$d\alpha^u = \alpha^u \log\alpha\, du \qquad (\alpha > 0). \tag{9}$$

As the reader can see we have simplified the notation. Confusion is not possible because of Cauchy's Rule of Invariance. For example, letting $\phi(x) = u(x) + v(x)$, the unsimplified version of (3) would be $d\phi(c;h) = du(c;h) + dv(c;h)$.

Similar results hold if U and V are matrix functions, and A is a matrix of real constants:

$$dA = 0 \tag{10}$$

$$d(\alpha U) = \alpha dU \tag{11}$$

$$d(U+V) = dU + dV \tag{12}$$

$$d(UV) = (dU)V + UdV\,. \tag{13}$$

For the Kronecker product and Hadamard product the analogue of (13) holds:

$$d(U \otimes V) = (dU) \otimes V + U \otimes dV \tag{14}$$

$$d(U * V) = (dU) * V + U * dV . \tag{15}$$

Finally we have

$$dU' = (dU)' \tag{16}$$

$$d\,\mathrm{vec}\,U = \mathrm{vec}\,dU \tag{17}$$

$$d\,\mathrm{tr}\,U = \mathrm{tr}\,dU . \tag{18}$$

We shall prove some of these results. For example, to prove (4), let $\phi(x) = u(x) + v(x)$. Then

$$\begin{aligned} d\phi(x;h) &= \sum_j h_j D_j \phi(x) = \sum_j h_j [D_j u(x) + D_j v(x)] \\ &= \sum_j h_j D_j u(x) + \sum_j h_j D_j v(x) = du(x;h) + dv(x;h). \end{aligned}$$

As a second example, let us prove (13). Using only (3) and (4), we have

$$\begin{aligned} \{d(UV)\}_{ij} &= d(UV)_{ij} \\ &= d \sum_k u_{ik} v_{kj} = \sum_k d(u_{ik} v_{kj}) \\ &= \sum_k \{(du_{ik}) v_{kj} + u_{ik} dv_{kj}\} \\ &= \sum_k (du_{ik}) v_{kj} + \sum_k u_{ik} dv_{kj} \\ &= \{(dU)V\}_{ij} + \{UdV\}_{ij} . \end{aligned}$$

Hence (13) follows. The Kronecker and Hadamard analogues (14) and (15) follow in a similar way.

10. Applications to simple, Hadamard, and Kronecker products

Let us now apply the theory developed in this paper to obtain the Jacobian matrices of the simple, Hadamard, and Kronecker product, respectively. These results correspond to Theorems 2-4 of Bentler and Lee (1978). The reader will see that the approach via differentials and the First Identification Theorem is elegant, simple, and straightforward.

Before stating Theorem 9 we offer a simple proof of the following well-known result.

Lemma 1. [Roth (1934), Neudecker (1969)]. For any three matrices A, B and C such that the matrix product ABC is defined,

$$\text{vec } ABC = (C' \otimes A) \text{ vec } B. \tag{1}$$

Proof. Assume that B has s columns denoted $b_1, b_2, \ldots, b_s$. Similarly, let the columns of the s×s identity matrix I_s be denoted $u_1, u_2, \ldots, u_s$. Then we can write $B = \sum_{j=1}^{s} b_j u_j'$, so that

$$\begin{aligned}\text{vec } ABC &= \text{vec} \sum_{j=1}^{s} Ab_j u_j' C = \sum_{j=1}^{s} \text{vec } (Ab_j)(C'u_j)' \\ &= \sum_{j=1}^{s} (C'u_j \otimes Ab_j) = (C' \otimes A) \sum_{j=1}^{s} (u_j \otimes b_j) \\ &= (C' \otimes A) \sum_{j=1}^{s} \text{vec } b_j u_j' = (C' \otimes A) \text{ vec } B. \quad \| \end{aligned} \tag{2}$$

Theorem 9 (Simple product). Let $U : S \to \mathbb{R}^{m \times r}$ and $V : S \to \mathbb{R}^{r \times p}$ be two matrix functions defined and differentiable on an open set S in $\mathbb{R}^{n \times q}$. Then the simple product UV is differentiable on S and the Jacobian matrix is the mp×nq matrix

$$\frac{\partial \text{ vec } UV}{\partial (\text{vec } X)'} = (V' \otimes I_m) \frac{\partial \text{ vec } U}{\partial (\text{vec } X)'} + (I_p \otimes U) \frac{\partial \text{ vec } V}{\partial (\text{vec } X)'} . \tag{3}$$

Proof. We take differentials,

$$d(UV) = (dU)V + U(dV) . \tag{4}$$

Vectorizing (4) we obtain, using Lemma 1,

$$\begin{aligned} \operatorname{vec} d(UV) &= \operatorname{vec} (dU)V + \operatorname{vec} U(dV) \\ &= (V' \otimes I_m) \operatorname{vec} dU + (I_p \otimes U) \operatorname{vec} dV \\ &= (V' \otimes I_m) \frac{\partial \operatorname{vec} U}{\partial (\operatorname{vec} X)'} \operatorname{vec} dX + (I_p \otimes U) \frac{\partial \operatorname{vec} V}{\partial (\operatorname{vec} X)'} \operatorname{vec} dX \\ &= \left\{(V' \otimes I_m) \frac{\partial \operatorname{vec} U}{\partial (\operatorname{vec} X)'} + (I_p \otimes U) \frac{\partial \operatorname{vec} V}{\partial (\operatorname{vec} X)'}\right\} \operatorname{vec} dX. \end{aligned} \tag{5}$$

The result now follows from the First Identification Theorem. ||

To prove the corresponding result for the Hadamard product we need the following useful, if simple, result.

Lemma 2. For any two matrices A and B of the same order,

$$\operatorname{vec} (A * B) = \hat{A} \operatorname{vec} B = \hat{B} \operatorname{vec} A, \tag{6}$$

where $\hat{A}$ denotes the diagonal matrix displaying the elements of vec A along its diagonal.

Proof. Obvious. ||

Theorem 10 (Hadamard product). Let $U : S \to \mathbb{R}^{m\times p}$ and $V : S \to \mathbb{R}^{m\times p}$ be two matrix functions defined and differentiable on an open set S in $\mathbb{R}^{n\times q}$. Then the Hadamard product $U * V$ is differentiable on S and the Jacobian matrix is the $mp\times nq$ matrix

$$\frac{\partial \operatorname{vec} (U * V)}{\partial (\operatorname{vec} X)'} = \hat{V} \frac{\partial \operatorname{vec} U}{\partial (\operatorname{vec} X)'} + \hat{U} \frac{\partial \operatorname{vec} V}{\partial (\operatorname{vec} X)'} . \tag{7}$$

Proof. We take differentials,

$$d(U * V) = (dU) * V + U * (dV), \tag{8}$$

and vectorize, using Lemma 2,

$$\begin{aligned} \text{vec } d(U * V) &= \text{vec } \big((dU) * V\big) + \text{vec } \big(U * (dV)\big) \\ &= \hat{V} \text{ vec } dU + \hat{U} \text{ vec } dV \\ &= \hat{V} \frac{\partial \text{ vec } U}{\partial (\text{vec } X)'} \text{ vec } dX + \hat{U} \frac{\partial \text{ vec } V}{\partial (\text{vec } X)'} \text{ vec } dX \\ &= \left\{\hat{V} \frac{\partial \text{ vec } U}{\partial (\text{vec } X)'} + \hat{U} \frac{\partial \text{ vec } V}{\partial (\text{vec } X)'}\right\} \text{ vec } dX. \end{aligned} \tag{9}$$

Again, the result now follows from the First Identification Theorem. ||

The result for the Kronecker product is more difficult, but not as difficult as Bentler and Lee (1978) appear to think when they precede their Theorem 4 with six lemmas. For us one lemma suffices.

Lemma 3. [Neudecker and Wansbeek (1983)]. For any $m \times p$ matrix A and $r \times s$ matrix B,

$$\text{vec } (A \otimes B) = (I_p \otimes K_{sm} \otimes I_r)(\text{vec } A \otimes \text{vec } B). \tag{10}$$

Proof. Let a_i $(i = 1,\ldots,p)$ and b_j $(j = 1,\ldots,s)$ denote the columns of A and B, respectively. Also, let e_i $(i = 1,\ldots,p)$ and u_j $(j = 1,\ldots,s)$ denote the columns of I_p and I_s, respectively. Then we can write A and B as

$$A = \sum_{i=1}^{p} a_i e_i' , \quad B = \sum_{j=1}^{s} b_j u_j' , \tag{11}$$

and we obtain

$$\text{vec}\,(A \otimes B) = \sum_{i=1}^{p}\sum_{j=1}^{s} \text{vec}\,(a_i e_i' \otimes b_j u_j')$$

$$= \sum_{ij} \text{vec}\,(a_i \otimes b_j)(e_i \otimes u_j)' = \sum_{ij} (e_i \otimes u_j \otimes a_i \otimes b_j)$$

$$= \sum_{ij} (I_p \otimes K_{sm} \otimes I_r)(e_i \otimes a_i \otimes u_j \otimes b_j)$$

$$= (I_p \otimes K_{sm} \otimes I_r)\{(\sum_i \text{vec}\, a_i e_i') \otimes (\sum_j \text{vec}\, b_j u_j')\}$$

$$= (I_p \otimes K_{sm} \otimes I_r)(\text{vec}\, A \otimes \text{vec}\, B). \quad || \qquad (12)$$

<u>Theorem 11. (Kronecker product)</u>. Let $U : S \to \mathbb{R}^{m\times p}$ and $V : S \to \mathbb{R}^{r\times s}$ be two matrix functions defined and differentiable on an open set S in $\mathbb{R}^{n\times q}$. Then the Kronecker product $U \otimes V$ is differentiable on S and the Jacobian matrix is the $mprs\times nq$ matrix

$$\frac{\partial\, \text{vec}\,(U \otimes V)}{\partial (\text{vec}\, X)'} = (I_p \otimes G) \frac{\partial\, \text{vec}\, U}{\partial (\text{vec}\, X)'} + (H \otimes I_r) \frac{\partial\, \text{vec}\, V}{\partial (\text{vec}\, X)'}, \qquad (13)$$

with

$$G = (K_{sm} \otimes I_r)(I_m \otimes \text{vec}\, V), \quad H = (I_p \otimes K_{sm})(\text{vec}\, U \otimes I_s). \qquad (14)$$

<u>Proof</u>. We take differentials,

$$d(U \otimes V) = (dU) \otimes V + U \otimes (dV), \qquad (15)$$

and vectorize, using Lemma 3,

$$\text{vec}\, d(U \otimes V) = \text{vec}\, \big((dU) \otimes V\big) + \text{vec}\,(U \otimes dV)$$

$$= (I_p \otimes K_{sm} \otimes I_r)(\text{vec}\, dU \otimes \text{vec}\, V + \text{vec}\, U \otimes \text{vec}\, dV)$$

$$= (I_p \otimes K_{sm} \otimes I_r)\{(I_{mp} \otimes \text{vec}\, V)\, \text{vec}\, dU + (\text{vec}\, U \otimes I_{rs})\text{vec}\, dV\}$$

$$= (I_p \otimes G)\, \text{vec}\, dU + (H \otimes I_r)\, \text{vec}\, dV$$

$$= \{(I_p \otimes G) \frac{\partial\, \text{vec}\, U}{\partial (\text{vec}\, X)'} + (H \otimes I_r) \frac{\partial\, \text{vec}\, V}{\partial (\text{vec}\, X)'}\}\, \text{vec}\, dX. \qquad (16)$$

The result then follows from the First Identification Theorem. ||

11. Applications to functions of complicated matrix products

The number of applications is, of course, boundless. Let us consider one final example, the function of five matrices $\operatorname{tr}[(A \otimes B)\ C\ (D * E)]^{-1}$. Bentler and Lee (1978) claim that functions such as these arise in various disciplines including biometrics and mathematical psychology. We wish to differentiate this function with respect to each of the elements of the various matrices. Bentler and Lee mention this problem, but they do not solve it.

We first prove Theorem 12.

Theorem 12. Let $U : S \to \mathbb{R}^{m\times r}$, $V : S \to \mathbb{R}^{r\times p}$, and $W : S \to \mathbb{R}^{p\times m}$ be three matrix functions defined and differentiable on an open set S in $\mathbb{R}^{n\times q}$. Let $F : S \to \mathbb{R}^{m\times m}$ be defined by $F(X) = U(X)V(X)W(X)$. Then, at every point X in S where the $m\times m$ matrix $F(X)$ is nonsingular,

$$\frac{\partial \operatorname{tr} F^{-1}}{\partial (\operatorname{vec} X)'} = -\left(\operatorname{vec} (F')^{-2}W'V'\right)'\left(\partial \operatorname{vec} U/\partial(\operatorname{vec} X)'\right)$$

$$-\left(\operatorname{vec} U'(F')^{-2}W'\right)'\left(\partial \operatorname{vec} V/\partial(\operatorname{vec} X)'\right)$$

$$-\left(\operatorname{vec} V'U'(F')^{-2}\right)'\left(\partial \operatorname{vec} W/\partial(\operatorname{vec} X)'\right). \qquad (1)$$

Proof. We have

$$d\operatorname{tr} F^{-1} = \operatorname{tr} dF^{-1} = -\operatorname{tr} F^{-1}(dF)F^{-1} = -\operatorname{tr} F^{-2}dF, \qquad (2)$$

and further

$$dF = d(UVW) = (dU)VW + U(dV)W + UV(dW). \qquad (3)$$

Inserting (3) in (2) and rearranging gives

$$d\,\mathrm{tr}\,F^{-1} = -\mathrm{tr}\,VWF^{-2}dU - \mathrm{tr}\,WF^{-2}UdV - \mathrm{tr}\,F^{-2}UVdW$$

$$= -\left(\mathrm{vec}\ (F')^{-2}W'V'\right)'\,\mathrm{vec}\,dU - \left(\mathrm{vec}\ U'(F')^{-2}W'\right)'\,\mathrm{vec}\,dV$$

$$-\left(\mathrm{vec}\ V'U'(F')^{-2}\right)'\ \mathrm{vec}\ dW, \qquad (4)$$

making use of the well-known relationship $\mathrm{tr}\,A'B = (\mathrm{vec}\ A)'\,\mathrm{vec}\,B$.

Now, since $\mathrm{vec}\ dU = (\partial\,\mathrm{vec}\,U/\partial(\mathrm{vec}\ X)')'\,\mathrm{vec}\,dX$ and similarly for $\mathrm{vec}\,dV$ and $\mathrm{vec}\,dW$, the result follows from the First Identification Theorem. ||

We now have all the ingredients to prove Theorem 13.

<u>Theorem 13</u>. Let $A : S \to \mathbb{R}^{\ell\times r}$, $B : S \to \mathbb{R}^{m\times s}$, $C : S \to \mathbb{R}^{rs\times p}$, $D : S \to \mathbb{R}^{p\times \ell m}$, and $E : S \to \mathbb{R}^{p\times \ell m}$ be five matrix functions defined and differentiable on an open set S in $\mathbb{R}^{n\times q}$. Let $F : S \to \mathbb{R}^{\ell m\times \ell m}$ be defined by

$$F(X) = \left(A(X) \otimes B(X)\right)\ C(X)\ \left(D(X) * E(X)\right). \qquad (5)$$

Then, at every point X in S where the $\ell m\times\ell m$ matrix $F(X)$ is nonsingular,

$$\frac{\partial\,\mathrm{tr}\,F^{-1}}{\partial(\mathrm{vec}\ X)'} = -\left(\mathrm{vec}\ (F')^{-2}W'C'\right)'(I_r \otimes G)\left(\partial\,\mathrm{vec}\,A/\partial(\mathrm{vec}\ X)'\right)$$
$$-\left(\mathrm{vec}\ (F')^{-2}W'C'\right)'(H \otimes I_m)\left(\partial\,\mathrm{vec}\,B/\partial(\mathrm{vec}\ X)'\right)$$
$$-\left(\mathrm{vec}\ U'(F')^{-2}W'\right)'\left(\partial\,\mathrm{vec}\,C/\partial(\mathrm{vec}\ X)'\right)$$
$$-\left(\mathrm{vec}\ C'U'(F')^{-2}\right)'\hat{E}\left(\partial\,\mathrm{vec}\,D/\partial(\mathrm{vec}\ X)'\right)$$
$$-\left(\mathrm{vec}\ C'U'(F')^{-2}\right)'\hat{D}\left(\partial\,\mathrm{vec}\,E/\partial(\mathrm{vec}\ X)'\right), \qquad (6)$$

where $U = A \otimes B$, $W = D * E$, and

$$G = (K_{s\ell} \otimes I_m)(I_\ell \otimes \mathrm{vec}\ B),\quad H = (I_r \otimes K_{s\ell})(\mathrm{vec}\ A \otimes I_s). \qquad (7)$$

<u>Proof</u>. Immediate from Theorems 10-12. ||

In fact, Theorem 13 answers a more general question than the one posed in the beginning of this section. As a special case, suppose that the matrices B, C, D, and E are functionally independent from A, i.e. $\partial\,\mathrm{vec}\,B/\partial(\mathrm{vec}\,A)' = O$ and similarly for C, D, and E. Then we obtain from (6) the gradient vector

$$\frac{\partial\,\mathrm{tr}\,F^{-1}}{\partial(\mathrm{vec}\,A)'} = -\left(\mathrm{vec}\,(F')^{-2}W'C'\right)'(I_r \otimes G). \tag{8}$$

The other four gradients are, of course, determined in the same way.

Appendix: A conversion table

The reader who wishes to compare our results with those of Bentler and Lee (1978) will notice the difference in notation of the two papers. The following conversion table may thus prove useful.

Bentler-Lee	Magnus-Neudecker
I^n	I_n
$\bar{X}$	$(\mathrm{vec}\ X')'$
$D_{\bar{X}}$	$\hat{X}$
E^{mn}	K_{mn}
$\frac{\partial Y}{\partial Z}$	$\left(\frac{\partial\ \mathrm{vec}\ Y'}{\partial(\mathrm{vec}\ Z')'}\right)'$

References

Balestra, P. La dérivation matricielle. Paris: Sirey, 1976.

Bentler, P.M., and Lee, S-Y. Matrix derivatives with chain rule and rules for simple, Hadamard, and Kronecker products. Journal of Mathematical Psychology, 1978, 17, 255-262.

Dwyer, P.S. Some applications of matrix derivatives in multivariate analysis. Journal of the American Statistical Association, 1967, 62, 607-625.

Dwyer, P.S., and MacPhail, M.S. Symbolic matrix derivatives. Annals of Mathematical Statistics, 1948, 19, 517-534.

MacRae, E.C. Matrix derivatives with an application to an adaptive linear decision problem. The Annals of Statistics, 1974, 2, 337-346.

Magnus, J.R., and Neudecker, H. The commutation matrix: some properties and applications. The Annals of Statistics, 1979, 7, 381-394.

McDonald, R.P., and Swaminathan, H. A simple matrix calculus with applications to multivariate analysis. General Systems, 1973, 18, 37-54.

Nel, D.G. On matrix differentiation in statistics. South African Statistical Journal, 1980, 14, 137-193.

Neudecker, H. Some theorems on matrix differentiation with special reference to Kronecker matrix products. Journal of the American Statistical Association, 1969, 64, 953-963.

Neudecker, H., and Wansbeek, T.J. Some results on commutation matrices, with statistical applications. The Canadian Journal of Statistics, 1983, 11, 221-231.

Polasek, W. Local resistence in polynomial regression and restricted least squares, in: Statistics and Probability (J. Mogyorodi, I. Vincze, and W. Wertz, Eds.). Dordrecht, The Netherlands: Reidel, 1984.

Pollock, D.S.G. Tensor products and matrix differential calculus (Revised version). Discussion paper No. 86, Department of Economics, Queen Mary College, London, 1984.

Rogers, G.S. Matrix derivatives. New York: Marcel Dekker, 1980.

Roth, W.E. On direct product matrices. Bulletin of the American Mathematical Society, 1934, 40, 461-468.