A PRIMER ON DIFFERENTIAL CALCULUS
FOR VECTORS AND MATRICES

S. R. Searle and H. V. Henderson[1/]

Biometrics Unit, Cornell University, Ithaca, New York

BU-1047-MB January, 1993

Abstract

A variety of authors have defined differential operators for vectors and matrices in different ways. Definitions and notation that constitute a smooth transition from the differential calculus of scalars to that of matrices are set out in this paper.

1. INTRODUCTION

The differential calculus of scalar variables dates back to Newton (1642-1727); and is obviously well known. Application to vectors and matrices is more recent, e.g., Turnbull [8]; and the decade of the 1980's has spawned a lengthy review by Nel [5] and a number of books on the subject, such as Rogers [6], Graham [3] and Magnus and Neudecker [4]. The latter is of particular interest to statisticians, as are numerous research papers listed therein that have appeared in the last twenty years or so. Most of these, of course, deal with advanced aspects of the subject. They also embody a variety of definitions and notations which Magnus and Neudecker ([4], pp. 171-5) critically discuss and evaluate. The purpose of this paper is to provide an introduction to what we believe is the best of these notations, doing so in a manner that provides a smooth transition from the differential calculus of scalar variables to that of matrix variables. It is also a notation which holds up for what we think are useful vector and matrix extensions of the product rule and the chain rule of scalar differentiation – and not all notations found in the literature have this feature.

[1/]Visiting Fellow, summer 1989, from the Statistics Section, Ruakura Agricultural Center, Hamilton, New Zealand, with supplementary support from the Mathematical Sciences Institute of Cornell University and from the New Zealand Mathematical Society.

Starting with x and y (a differentiable function of x) as scalars, dy/dx is a scalar too, and there is no opportunity for doubt as to what is meant by dy/dx. (Since this is largely an expository paper, it conveniently by-passes details of the mathematical foundations of differential calculus so that, for example, any regularity conditions necessary for the operations we discuss are assumed to be in place.) In contrast, when **x** and **y** are vectors (with elements of **y** being differentiable functions of elements of **x**), we have available the partial derivative of every element of **y** with respect to every element of **x**. This immediately poses two questions: (i) What is a convenient way of arraying these partial derivatives – as a row vector, as a column vector, or as a matrix, and if the latter of what order? (ii) What is a good symbol for this array: $\partial\mathbf{y}/\partial\mathbf{x}$, $\partial\mathbf{y}/\partial\mathbf{x}'$, $\partial\mathbf{y}'/\partial\mathbf{x}$ or $\partial\mathbf{y}'/\partial\mathbf{x}'$? We answer these questions by settling for the array being a matrix, of order $r \times \rho$ for $\mathbf{y}_{r\times 1}$ and $\mathbf{x}_{\rho\times 1}$, and we choose $\partial\mathbf{y}/\partial\mathbf{x}'$ to represent this matrix, as in (20).

Extension to matrices **Y** and **X** involves even more partial derivatives: of every element of **Y** with respect to every element of **X**. And here we use the vec notation: vec **X** is a column vector with the columns of **X** set one beneath the other, in order, so that for $\mathbf{X}_{\rho\times\gamma}$, the vector vec **X** has order $\rho\gamma \times 1$. Then, for $\mathbf{Y}_{r\times c}$ there are $rc\rho\gamma$ partial derivatives $\partial y_{ij}/\partial x_{pq}$ for $i = 1, \cdots, r$, $j = 1, \cdots, c$, $p = 1, \cdots, \rho$ and $q = 1, \cdots, \gamma$; and they are arrayed in a matrix

$$\frac{\partial(\text{vec }\mathbf{Y})}{\partial(\text{vec }\mathbf{X})'} \text{ of order } rc \times \rho\gamma ,$$

governed by exactly the same definition as is given for $\partial\mathbf{y}/\partial\mathbf{x}'$ in the preceding paragraph.

2. SCALAR RESULTS

We begin with a summary of results in scalar calculus which are convenient for developing analogous results in matrix algebra.

Differentiation For y being a function of x, f(x) say, dy/dx represents differentiating y with respect to x. Thus for

$$y = 3x^5 + e^{2x}, \quad \frac{dy}{dx} = 15x^4 + 2e^{2x} . \tag{1}$$

Product rule With u and v both being functions of x, then for

$$y = uv \quad \text{we have} \quad \frac{dy}{dx} = \frac{du}{dx}v + u\frac{dv}{dx}\,. \tag{2}$$

Proof is available in Crowell and Slesnick ([1], p. 59), a source that is hereafter denoted CS.

The differential For y being the function f(x) of x, dy/dx is often denoted f′(x):

$$\frac{dy}{dx} = f'(x)\,. \tag{3}$$

f′(x) is a function of x; an example is (1). And for any particular value of x, f′(x) is the slope of the curve y = f(x) at the point (x,y). Thus what is called the differential of y is defined as

$$dy = f'(x)dx\,. \tag{4}$$

The chain rule When y is a function of w, which is itself a function of x, then (see for example, CS, p. 69) for

$$y = f(w) \quad \text{and} \quad w = g(x), \quad \frac{dy}{dx} = \frac{dy}{dw}\frac{dw}{dx}\,. \tag{5}$$

And the differential from this is, akin to (4),

$$dy = \frac{dy}{dw}dw\,. \tag{6}$$

Partial differentiation Suppose y is a function $f(x_1, x_2, \cdots, x_\rho)$ of ρ x-variables. The differential of y with respect to x_1 say, assuming that all the other xs are held constant, is the partial derivative of y with respect to x_1 and is denoted $\partial y/\partial x_1$. Thus

$$\frac{\partial y}{\partial x_1} \text{ is } \left(\frac{dy}{dx_1} \text{ holding } x_2, x_3, \cdots, x_\rho \text{ constant}\right).$$

An extended chain rule Suppose y is a function $f(w_1, w_2, \cdots, w_k)$, where each w is a function of a single x. Then what we call an extended chain rule is

$$\frac{dy}{dx} = \sum_{i=1}^{k} \frac{\partial y}{\partial w_i}\frac{dw_i}{dx}\,. \tag{7}$$

This is clearly an extension of (5), but whereas (5) has dy/dw because there is only one w, equation (7) has $\partial y/\partial w_i$ for each w_i, $i = 1, \cdots, k$.

The total differential Based on (7), and akin to (6), what is called the total differential of y (e.g., Curtis, [2], p. 211) is

$$dy = \sum_{i=1}^{k} \frac{\partial y}{\partial w_i} dw_i\,. \tag{8}$$

Clearly, if k = 1, this reduces to (6).

Importance of the total differential

For scalars, the importance of the differential is that whenever we have an equation

$$dy = t\,dx$$

then, so long as t is not a function of dx, the identification theorem (Magnus and Neudecker, [4], p. 87) gives

$$\frac{dy}{dx} = t\ .$$

Although t must not be a function of dx it may or may not be a function of x. Similarly, from the total differential of (8), if

$$dy = \sum_{i=1}^{k} t_i dw_i$$

where t_i is not a function of any dw_i (but it can be a function of the w_is), then

$$\frac{\partial y}{\partial w_i} = t_i \quad \text{for} \quad i = 1, \cdots, k\ .$$

As a simple example, taking differentials of $y = 7x_1^2 + 3x_2^4 + e^{x_2}$ gives

$$dy = 14x_1 dx_1 + \left(12x_2^3 + e^{x_2}\right)dx_2$$

and so

$$\frac{\partial y}{\partial x_1} = 14x_1 \quad \text{and} \quad \frac{\partial y}{\partial x_2} = 12x_2^3 + e^{x_2}\ .$$

3. DIFFERENTIATION WITH RESPECT TO A VECTOR

Define $\mathbf{x} = [x_1 \ \ x_2 \ \ \cdots \ \ x_\rho]'$, as a column vector of order ρ, and consider differentiating functions of **x**, i.e., functions of $x_1, x_2, \cdots, x_\rho$. The functions can be scalars represented as $f(\mathbf{x})$, or vectors, $\mathbf{f}(\mathbf{x})$, or matrices $\mathbf{F}(\mathbf{x})$. With all of them we consider differentiation with respect to elements of **x**, which therefore involves more than one scalar x_i. Hence differentiation with respect to **x** is considered in terms of partial differentiation with respect to each element of **x**. Thus we always have an array of partial derivatives to deal with. This requires deciding how to display each array, and how to symbolize it, preferably with notation that has a mnemonic attribute for what the array is.

3.1 Differentiating a scalar function, f(x).

For scalar $y = f(\mathbf{x})$ we use $\partial y/\partial \mathbf{x}$ to represent the column vector of partial derivatives of y with respect to each element of $\mathbf{x}$:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_\rho} \end{bmatrix}. \tag{9}$$

In doing so, notice that in the symbol $\partial y/\partial \mathbf{x}$ the $\mathbf{x}$ is mnemonic for arraying the partial derivatives $\partial y/\partial x_1$, $\partial y/\partial x_2$, $\cdots$, $\partial y/\partial x_\rho$ as a column vector. Were we to array them as a row vector we would have $\mathbf{x}'$ in place of $\mathbf{x}$ and use the symbol

$$\frac{\partial y}{\partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_\rho} \end{bmatrix} = \left(\frac{\partial y}{\partial \mathbf{x}}\right)', \tag{10}$$

the transpose of (9).

3.2 Differentiating a vector with respect to a scalar

Differentiating a scalar with respect to a vector is shown in (9) and (10). The complement is differentiating a vector with respect to a scalar. This takes the forms

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_r}{\partial x} \end{bmatrix} \quad \text{and} \quad \frac{\partial \mathbf{y}'}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_r}{\partial x} \end{bmatrix} = \left(\frac{\partial \mathbf{y}}{\partial x}\right)'. \tag{11}$$

Example 1 Define $\mathbf{y} = \mathbf{a}'\mathbf{x} = \sum_{i=1}^{\rho} a_i x_i$. From (9)

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_\rho \end{bmatrix} = \mathbf{a}\,, \tag{12}$$

i.e.,

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}'\mathbf{x} = \mathbf{a}\,. \tag{13}$$

And because $\mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$,

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{a} = \mathbf{a}\,. \tag{14}$$

Then

$$\frac{\partial}{\partial \mathbf{x}'} \mathbf{a}'\mathbf{x} = \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{a}'\mathbf{x}\right)' = \mathbf{a}' = \frac{\partial}{\partial \mathbf{x}'} \mathbf{x}'\mathbf{a}\,. \tag{15}$$

The similarity of these results to the scalar $d(ax)/dx = a$ is clear; but what is different, and what is so important, is the need for deciding (defining) how to array the results and how to mnemonically and usefully symbolize them. In (13) and (14) we see that differentiation of $\mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$ with respect to the column vector $\mathbf{x}$ leads to the column vector $\mathbf{a}$; and in (15) differentiation with respect to the row vector $\mathbf{x}'$ leads to the row vector $\mathbf{a}'$.

Example 2 (1st time)

Define $$y = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{\rho} a_{ii} x_i^2 + \sum_{i \neq j = 1}^{\rho} a_{ij}\, x_i\, x_j\,.$$

Then $$\frac{\partial y}{\partial x_i} = 2a_{ii}x_i + \sum_{\substack{j=1 \\ j \neq i}}^{\rho} (a_{ij} + a_{ji}) x_j\,.$$

This can be rewritten as

$$\begin{aligned} \frac{\partial y}{\partial x_i} &= a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ii}x_i + \cdots + a_{i\rho}x_\rho + (a_{1i}x_i + a_{2i}x_2 + \cdots + a_{ii}x_i + \cdots + a_{\rho i}x_\rho) \\ &= \boldsymbol{\alpha}_i'\mathbf{x} + \mathbf{a}_i'\mathbf{x} \end{aligned} \tag{16}$$

where $\boldsymbol{\alpha}_i'$ and $\mathbf{a}_i$ are the i'th row and column, respectively, of **A**. Assembling (16) for all x_i, $i = 1, \cdots, \rho$ gives

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \dfrac{\partial y}{\partial x_1} \\ \vdots \\ \dfrac{\partial y}{\partial x_\rho} \end{bmatrix} = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x} \tag{17}$$

and when **A** is taken as symmetric, which it usually is,

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x} \quad \textbf{for} \quad \mathbf{A} = \mathbf{A}' \, . \tag{18}$$

3.3 **Differentiating a vector function, f(x).**

Suppose $\mathbf{y}'$ is a row vector of order r which we represent as

$$\mathbf{y}' = \left\{_r \, y_j\right\}_{j=1}^{r} \quad \text{with} \quad y_j = f_j(x) \, ,$$

where the r inside the braces stands for "row" (meaning that the elements y_j are arrayed as a row vector) and has nothing to do with the r outside the braces which indicates that there are r elements y_j. Differentiating $\mathbf{y}'$ with respect to $\mathbf{x}$ (of order ρ) is defined as taking the partial derivative of every element of $\mathbf{y}'$ with respect to every element of $\mathbf{x}$. This is a total of $r\rho$ partial derivatives, for which we have to settle on an array and a notation. In view of the vector array of $\partial y/\partial \mathbf{x}$ given in (9) for scalar y, and exemplified in (12), the natural array is the r columns $\partial y_j/\partial \mathbf{x}$ for $j = 1, \cdots, r$, set out as an $\rho \times r$ matrix. Thus for $\mathbf{y}_{r\times 1}$ and $\mathbf{x}_{\rho\times 1}$

$$\frac{\partial \mathbf{y}'}{\partial \mathbf{x}} = \left[\frac{\partial y_1}{\partial \mathbf{x}} \cdots \frac{\partial y_r}{\partial \mathbf{x}}\right] = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \dfrac{\partial y_2}{\partial x_1} & \cdots & \dfrac{\partial y_r}{\partial x_1} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial y_1}{\partial x_\rho} & \dfrac{\partial y_2}{\partial x_\rho} & \cdots & \dfrac{\partial y_r}{\partial x_\rho} \end{bmatrix}, \text{ of order } \rho \times r \tag{19}$$

which in a messy but self-contained notation we could write as

$$\frac{\partial(\mathbf{y}')_{1\times r}}{\partial \mathbf{x}_{\rho\times 1}} = \left\{ \frac{\partial y_j}{\partial x_i} \right\}_{m\; i=1\; j=1}^{\;\;\rho\;\;\; r} \quad \text{of order } \rho \times r \,.$$

Note that it has ρ rows as does **x** and r columns as does $\mathbf{y}'$. Magnus and Neudecker ([4], p. 87) call this the gradient matrix. Its ij'th element (in row i and column j) is $\partial y_j/\partial x_i$.

Note that if y is a scalar then $\partial\mathbf{y}'/\partial\mathbf{x}$ is just $\partial y/\partial\mathbf{x}$ and is the first column of (19), and this is (9), which it should be. And if **x** is a scalar then $\partial\mathbf{y}'/\partial\mathbf{x}$ is simply $\partial\mathbf{y}'/\partial x$ and is the first row of (19), and this is (11), as it should be.

In the symbol $\partial\mathbf{y}'/\partial\mathbf{x}$ representing the matrix of (19), $\mathbf{y}'$ indicates that the partial derivatives $\partial y_j/\partial x_i$ for $j = 1, \cdots, r$ are arrayed as a row, for each x_i, so that the rows of (19) are all based on the row vector $\mathbf{y}'$. Hence $\partial\mathbf{y}'/\partial\mathbf{x}$ has r columns, corresponding to the r elements of **y**. Similarly the column symbol **x** in $\partial\mathbf{y}'/\partial\mathbf{x}$ indicates that the terms $\partial y_j/\partial x_i$ for $i = 1, \cdots, \rho$ are arranged, for each y_j, as a column, and so $\partial\mathbf{y}'/\partial\mathbf{x}$ has ρ rows corresponding to the ρ elements of **x**. Hence $\partial\mathbf{y}'/\partial\mathbf{x}$ is, as seen in (19), a matrix of order $\rho \times r$.

Naturally, one can also transpose (19) and have

$$\frac{\partial\mathbf{y}}{\partial\mathbf{x}'} = \left(\frac{\partial\mathbf{y}'}{\partial\mathbf{x}}\right)' = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_\rho} \\ \cdots & & \cdots \\ \frac{\partial y_r}{\partial x_1} & \cdots & \frac{\partial y_r}{\partial x_\rho} \end{bmatrix} . \tag{20}$$

with element $\partial y_i/\partial x_j$ in the j'th row and i'th column.

Magnus and Neudecker ([4], p.87) call this the Jacobian matrix; in being the transpose of (18) it is, of course, the transpose of their gradient matrix. And when $r = \rho$, and both (18) and (20) are square, the absolute value of the determinant of either matrix is called the Jacobian (determinant). In statistics, when **x** and **y** are vectors of random variables, this Jacobian plays an important role in deriving the probability density function of **y** from that of **x**.

No matter which of (19) or (20) one uses, each is the transpose of the other, and whichever is the most useful will depend upon context. What is important is the manner of arraying the terms $\partial y_j/\partial x_i$ in $\partial \mathbf{y}'/\partial \mathbf{x}$ and $\partial \mathbf{y}/\partial \mathbf{x}'$. The mnemonic nature of $\partial \mathbf{y}'/\partial \mathbf{x}$ has been described, and that description applies in comparable fashion to $\partial \mathbf{y}/\partial \mathbf{x}'$.

Note in passing that with these specific meanings for $\partial \mathbf{y}'/\partial \mathbf{x}$ and $\partial \mathbf{y}/\partial \mathbf{x}'$, there is no obvious meaning for $\partial \mathbf{y}/\partial \mathbf{x}$ or $\partial \mathbf{y}'/\partial \mathbf{x}'$. The convenience of using a symbol that has mnemonic value without the need for verbal description is apparent: e.g., $\partial \mathbf{y}'/\partial \mathbf{x}$ has $\partial \mathbf{y}'$ as a row, so to speak, and $\partial \mathbf{x}$ as a column.

Perhaps we could devise a definition for $\partial \mathbf{y}/\partial \mathbf{x}$ and $\partial \mathbf{y}'/\partial \mathbf{x}'$. One possibility would be

$$\frac{\partial \mathbf{y}'}{\partial \mathbf{x}'} = \left[\mathrm{vec}(\partial \mathbf{y}'/\partial \mathbf{x})\right]' = \left[\frac{\partial y_1}{\partial x_1}\ \frac{\partial y_1}{\partial x_2}\ \cdots\ \frac{\partial y_1}{\partial x_r}\ \frac{\partial y_2}{\partial x_1}\ \cdots\ \frac{\partial y_2}{\partial x_r}\ \cdots\ \cdots\ \frac{\partial y_c}{\partial x_1}\ \cdots\ \frac{\partial y_c}{\partial x_r}\right].$$

Similarly we would define $\partial \mathbf{y}/\partial \mathbf{x} = \mathrm{vec}(\partial \mathbf{y}'/\partial \mathbf{x}) \equiv (\partial \mathbf{y}'/\partial \mathbf{x}')'$.

Example 3

Define $\mathbf{y} = \mathbf{A}\mathbf{x}$ for $\mathbf{A}$ of order $r \times \rho$. Let $\boldsymbol{\alpha}_i'$ be the i'th row of $\mathbf{A}$. Applying (13) to $y_i = \boldsymbol{\alpha}_i'\mathbf{x}$, the i'th element of $\mathbf{y}$, gives

$$\frac{\partial y_i}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}\boldsymbol{\alpha}_i'\mathbf{x} = \boldsymbol{\alpha}_i\ .$$

Hence

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \left(\frac{\partial \mathbf{y}'}{\partial \mathbf{x}}\right)' = \left[\frac{\partial y_1}{\partial \mathbf{x}}\ \cdots\ \frac{\partial y_r}{\partial \mathbf{x}}\right]' = [\boldsymbol{\alpha}_1\ \cdots\ \boldsymbol{\alpha}_r]' = \mathbf{A}\ ;$$

i.e.,

$$\frac{\partial}{\partial \mathbf{x}'}\mathbf{A}\mathbf{x} = \mathbf{A}\ . \qquad (21)$$

Equivalently, from (20)

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{x}'} \\ \frac{\partial y_2}{\partial \mathbf{x}'} \\ \vdots \\ \frac{\partial y_r}{\partial \mathbf{x}'} \end{bmatrix} = \begin{bmatrix} \alpha_1' \\ \alpha_2' \\ \vdots \\ \alpha_r' \end{bmatrix} = \mathbf{A}$$

Likewise for $\mathbf{y}' = \mathbf{x}'\mathbf{B}$, with $\mathbf{b}_j$ being the j'th column of $\mathbf{B}$,

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{x}'\mathbf{B} = \mathbf{B}\,.$$

Example 4

Define $y = \mathbf{u}'\mathbf{v}$ where, to begin with, each element of $\mathbf{u}$ and of $\mathbf{v}$ is a function of the scalar x. Then using the product rule of (2),

$$\frac{dy}{dx} = \frac{d}{dx}\Sigma_i u_i v_i = \Sigma_i \frac{du_i}{dx} v_i + \Sigma_i u_i \frac{dv_i}{dx} \tag{22}$$

which leads to a direct extension of (2):

$$\frac{d(\mathbf{u}'\mathbf{v})}{dx} = \frac{d\mathbf{u}'}{dx}\mathbf{v} + \mathbf{u}'\frac{d\mathbf{v}}{dx}\,. \tag{23}$$

But now suppose each element of $\mathbf{u}$ and of $\mathbf{v}$ is a function of elements of the vector $\mathbf{x}$. Then (22) is altered only by having partial derivatives with respect to x in place of derivatives with respect to x. Thus (22) becomes

$$\frac{\partial y}{\partial \mathbf{x}} = \Sigma_i \frac{\partial u_i}{\partial \mathbf{x}} v_i + \Sigma_i u_i \frac{\partial v_i}{\partial \mathbf{x}}\,. \tag{24}$$

Each partial derivative term on the right-hand side of (24) is, by (9), a vector. Hence (24) is a weighted sum of such vectors which, by consideration of conformability, leads us to writing it as

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial(\mathbf{u}'\mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}'}{\partial \mathbf{x}}\mathbf{v} + \frac{\partial \mathbf{v}'}{\partial \mathbf{x}}\mathbf{u}\,. \tag{25}$$

Compared to (23) the reversal of symbols in the second term of (25) is to be noted; and this cannot be otherwise because (25) is a column vector and each product in (25) is a matrix pre-multiplying a column.

Example 2 (2nd time)

With (25) available, the differentiation of $\mathbf{x}'\mathbf{A}\mathbf{x}$ is easy (by using $\mathbf{x}'$ as $\mathbf{u}'$ and $\mathbf{A}\mathbf{x}$ as $\mathbf{v}$):

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{x}'\mathbf{A}\mathbf{x} = \frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\mathbf{A}\mathbf{x} + \frac{\partial (\mathbf{A}\mathbf{x})'}{\partial \mathbf{x}}\mathbf{x}$$

$$= \mathbf{I}\mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{I}\mathbf{x} = (\mathbf{A} + \mathbf{A}')\mathbf{x}, \text{ as in (17)} . \tag{26}$$

3.4 **A chain rule for vectors.**

Using the notation for differentiation with respect to a vector, namely $\partial y/\partial \mathbf{x}$ of (9), the extended chain rule (7) can be written in vector form:

$$\frac{dy}{dx} = \frac{\partial y}{\partial \mathbf{w}'}\,\frac{d\mathbf{w}}{dx} . \tag{27}$$

This in turn can be extended to the case when scalar x is replaced by vector **x**. For then (27) holds for every element of **x** and we have the row vector

$$\frac{\partial y}{\partial \mathbf{x}'} = \frac{\partial y}{\partial \mathbf{w}'}\,\frac{\partial \mathbf{w}}{\partial \mathbf{x}'} . \tag{28}$$

Extending this again, to where y is replaced by **y** gives the chain rule for vectors

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \frac{\partial \mathbf{y}}{\partial \mathbf{w}'}\,\frac{\partial \mathbf{w}}{\partial \mathbf{x}'} . \tag{29}$$

Notice, once more, how the mnemonic characteristic of our notation clearly indicates the nature of the different results: (27) is a scalar, being a row premultiplying a column, (28) is a row, being a row premultiplying a matrix and (29) is a matrix, the product of two matrices. In terms of its elements it is, for $\mathrm{y}_{r \times 1}$, $\mathrm{w}_{k \times 1}$ and $\mathrm{x}_{\rho \times 1}$,

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \left\{_{m}\ \sum_{s=1}^{k} \frac{\partial y_i}{\partial w_s}\,\frac{\partial w_s}{\partial x_j}\right\}_{i=1\ j=1}^{r\ \ \ \rho}, \text{ of order } r \times \rho,$$

similar to (20).

3.5 **Partial derivatives and differentials of vectors.**

When y is a scalar function of (elements of) $\mathbf{x}$ we get from (8) that the total differential of $\mathbf{y}$ with respect to $\mathbf{w}$ is

$$dy = \frac{\partial y}{\partial \mathbf{w}'}d\mathbf{w} . \tag{30}$$

And for the sake of convention using the letter $\mathbf{x}$ in place of $\mathbf{w}$ gives

$$dy = \frac{\partial y}{\partial \mathbf{x}'}d\mathbf{x} . \tag{31}$$

Moreover, this extends very naturally, to vector $\mathbf{y}$ rather than scalar y:

$$d\mathbf{y} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}'}d\mathbf{x} . \tag{32}$$

Thus whenever applying rules on differentials we arrive at an equation $d\mathbf{y} = \mathbf{A}d\mathbf{x}$ (provided $\mathbf{A}$ is not a function of dx, but it may or may not be a function of $\mathbf{x}$), the matrix of partial derivatives can be identified as

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \mathbf{A} . \tag{33}$$

Use of differentials in this way avoids the tedium of going back to matrix elements and assembling scalar results into matrices.

Example 2 (3rd time) $y = \mathbf{x}'\mathbf{A}\mathbf{x}$ leads to

$$dy = d(\mathbf{x}'\mathbf{A}\mathbf{x}) = d\mathbf{x}'(\mathbf{A}\mathbf{x}) + \mathbf{x}'[d(\mathbf{A}\mathbf{x})] = (\mathbf{A}\mathbf{x})'d\mathbf{x} + \mathbf{x}'\mathbf{A}d\mathbf{x} = \mathbf{x}'(\mathbf{A}' + \mathbf{A})d\mathbf{x} .$$

Therefore on using (33)

$$\frac{\partial}{\partial \mathbf{x}'}\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'(\mathbf{A}' + \mathbf{A})$$

and so we get (17) and (26):

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{x}'\mathbf{A}\mathbf{x} = [\mathbf{x}'(\mathbf{A}' + \mathbf{A})]' = (\mathbf{A} + \mathbf{A}')\mathbf{x} .$$

Example 3 (2nd time) For $\mathbf{y} = \mathbf{A}\mathbf{x}$, $d\mathbf{y} = d(\mathbf{A}\mathbf{x}) = \mathbf{A}d\mathbf{x}$, and so (33) gives

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \frac{\partial}{\partial \mathbf{x}'}\mathbf{A}\mathbf{x} = \mathbf{A}, \text{ as in (21)} .$$

Example 4 (2nd time)

$$d(\mathbf{u}'\mathbf{v}) = d(\mathbf{u}')\mathbf{v} + \mathbf{u}'(d\mathbf{v}) = \mathbf{v}'d\mathbf{u} + \mathbf{u}'d\mathbf{v}\ . \tag{34}$$

Now, when elements of **u** and **v** are functions of elements of **x**, we have from (32)

$$d\mathbf{u} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}'}d\mathbf{x} \quad \text{and} \quad d\mathbf{v} = \frac{\partial \mathbf{v}}{\partial \mathbf{x}'}d\mathbf{x}\ ,$$

so that substitution in (34) gives

$$d(\mathbf{u}'\mathbf{v}) = \left(\mathbf{v}'\frac{\partial \mathbf{u}}{\partial \mathbf{x}'} + \mathbf{u}'\frac{\partial \mathbf{v}}{\partial \mathbf{x}'}\right)d\mathbf{x}$$

from which (33) yields

$$\frac{\partial(\mathbf{u}'\mathbf{v})}{\partial \mathbf{x}'} = \mathbf{v}'\frac{\partial \mathbf{u}}{\partial \mathbf{x}'} + \mathbf{u}'\frac{\partial \mathbf{v}}{\partial \mathbf{x}'} \tag{35}$$

which is the transpose of (25).

4. DIFFERENTIATION WITH RESPECT TO A MATRIX

The meaning of differentiating a matrix $\mathbf{Y}_{r\times c}$ with respect to a matrix $\mathbf{X}_{\rho\times\gamma}$ is taken to be differentiating every element y_{ij} of **Y** with respect to every element x_{pq} of **X**. (This assumes, of course, that every y_{ij} is a constant or a function of some or all of the x_{pq}s.) We therefore have $rc\rho\gamma$ partial derivatives $\partial y_{ij}/\partial x_{pq}$. The problem is how to usefully array them in a manner analagous to the arrays already defined, such as the vectors $\partial y/\partial \mathbf{x}$ and $\partial y/\partial \mathbf{x}'$ in (9) and (10), and the matrices $\partial \mathbf{y}'/\partial \mathbf{x}$ and $\partial \mathbf{y}/\partial \mathbf{x}'$ in (19) and (20). Since in (20) we already have a meaning for $\partial \mathbf{y}/\partial \mathbf{x}'$, where **x** and **y** are vectors, we define the differentiation of **Y** with respect to **X** by vectorizing **X** and **Y**. Thus the differentiation of **Y** with respect to **X** is taken to be the $rc \times \rho\gamma$ matrix

$$\frac{\partial(\text{vec}\ \mathbf{Y})}{\partial(\text{vec}\ \mathbf{X})'}\ . \tag{36}$$

Of the limitless examples available we have selected just two to illustrate (36).

Example 5 For

$$\mathbf{Y} = \mathbf{AXB}\ ,$$

it is a standard result (e.g., Searle, [7], p. 333) that

$$\text{vec } \mathbf{Y} = (\mathbf{B}' \otimes \mathbf{A})\text{vec } \mathbf{X}$$

where $\otimes$ represents the direct (or Kronecker) product operation. Hence for differentiating $\mathbf{Y} = \mathbf{AXB}$ with respect to $\mathbf{X}$ we have

$$\frac{\partial(\text{vec } \mathbf{Y})}{\partial(\text{vec } \mathbf{X})'} = \mathbf{B}' \otimes \mathbf{A}\ , \tag{37}$$

where $\mathbf{A}$ and $\mathbf{B}$ are not functions of $\mathbf{X}$.

Example 6

Starting with

$$\mathbf{Y} = \mathbf{X}^{-1}$$

we obtain the derivative of $\mathbf{X}^{-1}$ with respect to $\mathbf{X}$ by the following steps.

$$\mathbf{XY} = \mathbf{I}$$

$$(\text{d}\mathbf{X})\mathbf{Y} + \mathbf{X}(\text{d}\mathbf{Y}) = \mathbf{0}$$

$$\text{d}\mathbf{Y} = -\mathbf{X}^{-1}(\text{d}\mathbf{X})\mathbf{Y} = -\mathbf{X}^{-1}(\text{d}\mathbf{X})\mathbf{X}^{-1}$$

$$\text{vec}(\text{d}\mathbf{Y}) = -(\mathbf{X}^{-1\prime} \otimes \mathbf{X}^{-1})\text{vec}(\text{d}\mathbf{X})$$

$$\text{d}(\text{vec } \mathbf{Y}) = -(\mathbf{X}^{-1\prime} \otimes \mathbf{X}^{-1})\text{d}(\text{vec } \mathbf{X})\ .$$

Therefore on using (33)

$$\frac{\partial[\text{vec}(\mathbf{X}^{-1})]}{\partial(\text{vec } \mathbf{X})'} = -(\mathbf{X}^{-1\prime} \otimes \mathbf{X}^{-1})\ . \tag{38}$$

Our definition in (36), of differentiating $\mathbf{Y}$ with respect to $\mathbf{X}$, involves differentiation of vectors. Therefore all the rules and notation of Section 3 can be applied to (36) directly. For example, if $\mathbf{Y} = \mathbf{F}_1(\mathbf{W})$ and $\mathbf{W} = \mathbf{F}_2(\mathbf{X})$ a chain rule analogy of (29) that can be thought of as a chain rule for matrices is

$$\frac{\partial(\text{vec } \mathbf{Y})}{\partial(\text{vec } \mathbf{X})'} = \frac{\partial(\text{vec } \mathbf{Y})}{\partial(\text{vec } \mathbf{W})'}\ \frac{\partial(\text{vec } \mathbf{W})}{\partial(\text{vec } \mathbf{X})'}\ . \tag{39}$$

5. SCALAR FUNCTIONS OF MATRICES

Although the preceding notation for differentiating by elements of vectors and matrices is clearly very useful in a variety of situations, there are occasions when staying with basic elements is easiest. Two simple examples are, for $\mathbf{X}$ of order $\rho \times \gamma$

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{X}) = \left\{_{m} \frac{\partial}{\partial x_{ij}} \Sigma_i x_{ii} \right\}_{i=1}^{\rho} {}_{j=1}^{\gamma} = \mathbf{I} \,. \tag{40}$$

and

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{X}'\mathbf{X}) = \frac{\partial}{\partial \mathbf{X}} \sum_{i,j} x_{ij}^2 = \left\{_{m} \frac{\partial}{\partial x_{pq}} \sum_{i,j} x_{ij}^2 \right\}_{p=1}^{\rho} {}_{q=1}^{\gamma} = 2\mathbf{X} \,. \tag{41}$$

Another is differentiating $|\mathbf{X}|$. We start with

$$|\mathbf{X}| = \Sigma_j x_{ij} c_{ij} \qquad \forall\ i$$

where c_{ij} is the co-factor of x_{ij} in $|\mathbf{X}|$, the signed minor of x_{ij}. Then, for elements of $\mathbf{X}$ being functionally independent

$$\frac{\partial |\mathbf{X}|}{\partial x_{ij}} = c_{ij} \,. \tag{42}$$

Therefore

$$\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = \left\{ \frac{\partial |\mathbf{X}|}{\partial x_{ij}} \right\} = \{c_{ij}\} = [\operatorname{adj}(\mathbf{X})]' = |\mathbf{X}|\mathbf{X}^{-1\prime} \,, \tag{43}$$

where $\operatorname{adj}(\mathbf{X})$ is the adjoint of $\mathbf{X}$.

When $\mathbf{X}$ is symmetric, (43) becomes

$$\begin{aligned} \frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} &= 2(\operatorname{adj} \mathbf{X}) - \operatorname{diag}\{c_{11}\ \ c_{22}\ \cdots\ c_{\rho\rho}\} \\ &= |\mathbf{X}|(2\mathbf{X}^{-1} - \operatorname{diag}\{x^{11}\ \ x^{22}\ \ldots\ x^{\rho\rho}\}) \end{aligned}$$

where $\operatorname{diag}\{a\ \ b\ \ c\}$ is a diagonal matrix of elements a, b and c, and x^{ii} is the i'th diagonal element of $\mathbf{X}^{-1}$.

When either or both of $\mathbf{X}$ and $\mathbf{Y}$ is a scalar the form of (36) not needed; e.g., for $\mathbf{Y}$ being scalar

$$\frac{\partial y}{\partial \mathbf{X}} = \left\{_{m} \frac{\partial y}{\partial x_{pq}} \right\}_{p=1}^{\rho} {}_{q=1}^{\gamma} \ ; \text{ and for } \mathbf{X} \text{ being scalar } \ \frac{\partial \mathbf{Y}}{\partial x} = \left\{_{m} \frac{\partial y_{ij}}{\partial x} \right\}_{i=1}^{r} {}_{j=1}^{c} \,.$$

6. OTHER NOTATIONS

An oft-seen notation for the derivative of one vector with respect to another, of $\mathbf{y}$ with respect to $\mathbf{x}$ is $\partial\mathbf{y}/\partial\mathbf{x}$. It seems such an obvious extension of $\partial y/\partial x$ for scalars. But it has what we feel is a serious deficiency that as a symbol it contains no indication of being a matrix, let alone any suggestion of what order that matrix is. In contrast, our (19) and (20) avoid this deficiency. Thus when $\frac{\partial\mathbf{y}'}{\partial\mathbf{x}}$ is $\frac{\partial(\mathbf{y}_{r\times 1})'}{\partial\mathbf{x}_{\rho\times 1}}$ it is a matrix of order $\rho \times r$, and its transpose, $\frac{\partial\mathbf{y}_{r\times 1}}{\partial(\mathbf{x}_{\rho\times 1})'}$, is $r \times \rho$. This feature of the notation is particularly helpful, in the chain rules of (27) – (29). Thus (29) is $\frac{\partial\mathbf{y}_{r\times 1}}{\partial(\mathbf{x}_{\rho\times 1})'} = \frac{\partial\mathbf{y}_{r\times 1}}{\partial(\mathbf{w}_{k\times 1})'}\frac{\partial\mathbf{w}_{k\times 1}}{\partial(\mathbf{x}_{\rho\times 1})'}$ is a product of matrices of order $r \times k$ and $k \times \rho$ and hence is $r \times \rho$. In contrast, Graham [3, equation (4.6)] would write (29) as $\frac{\partial\mathbf{y}}{\partial\mathbf{w}}\frac{\partial\mathbf{w}}{\partial\mathbf{x}}$ which, without an accompanying verbal description, contains no hint of what form of vector or matrix it is.

Rather than vectorizing matrices $\mathbf{X}$ and $\mathbf{Y}$ as a step towards defining the derivative of $\mathbf{Y}$ with respect to $\mathbf{X}$ as we do in (35), other authors have used a variety of definitions involving Kronecker products and/or vec-permutation matrices. The latter are used by Rogers ([6], Theorem 5.2) where, for example, for $\mathbf{X}$ of order 2×3, he writes

$$\frac{\partial\mathbf{X}}{\partial\mathbf{X}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} = \text{vec}(\mathbf{I}_2)\left[\text{vec}(\mathbf{I}_3)\right]' \tag{44}$$

which has rank one. Instead, because $\partial\mathbf{t}/\partial\mathbf{t}' = \mathbf{I}$ for any vector $\mathbf{t}$, we would have

$$\frac{\partial(\text{vec }\mathbf{X})}{\partial(\text{vec }\mathbf{X})'} = \mathbf{I}_6 \,, \tag{45}$$

an identity matrix of order equal to the number of elements in $\mathbf{X}$. A connection of (45) to (44) is that $\mathbf{I}_6$ in (45) is $\mathbf{I}_2 \otimes \mathbf{I}_3$. We feel (45) is conceptually easier and more useful than (44). Magnus and Neudecker ([4], pages 171-5) would appear to agree with us. Not only does (45) display the partial derivatives but it gives the Jacobian matrix for the identity transformation $\mathbf{y} = \mathbf{x}$.

Bibliography

[1] Crowell, R.H. and Slesnick, W.E. (1968). *Calculus with Analytic Geometry*, Norton, New York.

[2] Curtis, P.C., Jr. (1972). *Multivariate Calculus with Linear Algebra*, Wiley, New York.

[3] Graham, A. (1981). *Kronecker Products and Matrix Calculus*, Ellis Harwood, Chichester.

[4] Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Chichester.

[5] Nel, D.G. (1980). On matrix differentiation in statistics. *South African Statistical Journal* **14**, 137-193.

[6] Rogers, G.S. (1980). *Matrix Derivatives*, Marcel Dekker, New York.

[7] Searle, S.R. (1987). *Matrix Algebra Useful for Statistics*, John Wiley & Sons, New York.

[8] Turnbull, H.W. (1928). *Theory of Determinants, Matrices and Invariants*, London and Glasgow.