Matrix derivatives: Why and where did it go wrong?

Jan R. Magnus
Department of Econometrics & Data Science, Vrije Universiteit Amsterdam,
and Tinbergen Institute, Amsterdam, The Netherlands
email: jan@janmagnus.nl

March 6, 2024

1 Introduction

The modern theory of matrix calculus rests on two pillars: a correct definition of matrix derivative and the use of differentials. The necessity of a correct mathematical definition is obvious, but the use of differentials is also essential, because a differential does not alter the dimension of the matrix on which it operates. In this note I will only briefly touch on differentials, and I will concentrate on describing the historical path towards finding the correct definition of matrix derivative.

If we have a vector function $y = Ax$, where y has dimension $m \times 1$ and x has dimension $n \times 1$, then the derivative (Jacobian matrix) is an $m \times n$ matrix denoted by $Dy(x)$ or $\partial y/\partial x'$, such that column i contains the partial derivatives of the components of y with respect to x_i and row s contains the partial derivatives of y_s with respect to the components of x. Since we are working with vectors y and x, a natural one-dimensional ordering of their components exists:

$$y = (y_1, y_2, \ldots, y_m)', \qquad x = (x_1, x_2, \ldots, x_n)'.$$

There is no controversy about this definition.

The difficulty arises when we move from vectors to matrices. Now we have an $m \times p$ matrix function $Y = Y(X)$, where each element y_{st} of Y depends on an $n \times q$ matrix $X = (x_{ij})$. This produces $mnpq$ partial derivatives and the question arises how these should be ordered. The matrices Y and X have

a natural *two*-dimensional ordering of their elements, but no natural *one*-dimensional ordering. A one-dimensional ordering needs to be constructed, and this can be done in many ways of which $\operatorname{vec} Y$ and $\operatorname{vec} X$ have become standard. (The vector $\operatorname{vec} X$ contains the columns of X, one underneath the other, starting with the first column.)

It is understandable that one wishes to maintain the matrix structure of Y and X when ordering the $mnpq$ partial derivatives. Such an ordering is possible, but it does not lead to the correct definition of matrix derivative. The correct definition depends on a *one*-dimensional ordering of the matrices Y and X, that is, on $\operatorname{vec} Y$ and $\operatorname{vec} X$, because it is essential that each column of the derivative matrix contains the partial derivatives of all elements of Y with respect to a specific element of X and that each row contains the partial derivatives of a specific element of Y with respect to all elements of X.

This mathematical necessity is in conflict with the intuitive desire to maintain the matrix structures of Y and X, and I think that this explains why it has taken so long to arrive at the correct definition and why there still exists so much resistance against it.

2 The pioneers

Until around 1970, the usual way of tackling matrix calculus problems was to resort to scalar differential calculus. A function expressed compactly in terms of matrices would be expanded in scalar form, differentiated by scalar methods, and then reassembled in the desired matrix form. Thus, to obtain the derivative of the $n \times n$ matrix function $Y = X^{-1}$ one would evaluate the n^4 elements $\partial y_{st}/\partial x_{ij}$ and reassemble in some way. Nowadays, we would write $dX^{-1} = -X^{-1}(dX)X^{-1}$, then vectorize to obtain

$$d\operatorname{vec} Y = d\operatorname{vec} X^{-1} = -(X'^{-1} \otimes X^{-1})d\operatorname{vec} X,$$

and conclude that $DY(X) = -(X'^{-1} \otimes X^{-1})$ is the derivative.[1]

Perhaps it was Herbert Turnbull who undertook the first attempt to define a matrix derivative. For any square $n \times n$ matrix function Y of a square $n \times n$ matrix X, Turnbull (1928) introduces the operator Ω such that

$$(\Omega Y)_{ij} = \sum_{s=1}^{n} \frac{\partial y_{sj}}{\partial x_{si}}, \tag{1}$$

and obtains some results based on this definition. This is obviously a very special case and thus does not provide a general theory of matrix calculus.

[1]The vec-operator, Kronecker product $\otimes$, commutation matrix K, and duplication matrix D are defined and discussed in *Matrix Differential Calculus*, among others.

The real pioneers were Paul Dwyer and Jack Macphail. In their 1948 paper, they consider two special cases, namely the case where X is a scalar ($n = q = 1$) and the case where Y is a scalar ($m = p = 1$). If X is a scalar, say x, and we have a matrix function $Y = Y(x)$, then we are understandably tempted to define its 'derivative' as $\dot{Y}$ (produced in LaTeX by the command `\dot{Y}`) such that $\dot{Y}$ has the same dimension as Y, for example

$$Y = \begin{pmatrix} x & 2x^3 & 3x^{-4} \\ e^x & \sin x & \log x \end{pmatrix} \implies \dot{Y} = \begin{pmatrix} 1 & 6x^2 & -12x^{-5} \\ e^x & \cos x & x^{-1} \end{pmatrix},$$

where the example is taken from Dwyer and Macphail (1948), but the notation $\dot{Y}$ is my own, emphasizing the fact that this matrix contains all the partial derivatives but is *not* the derivative $DY(x)$. More generally, for any $m \times p$ matrix function $Y = Y(x)$ of a scalar x, we have

$$\dot{Y}(x) = \sum_{s=1}^{m} \sum_{t=1}^{p} \frac{dy_{st}}{dx} E_{st}, \tag{2}$$

where E_{st} is a matrix (of appropriate order) containing one in the st-th position and zeros elsewhere.

Alternatively, if Y is a scalar function, say y, of a matrix $X = (x_{ij})$, then we may be tempted to define its 'derivative' as $\breve{y}$, for example

$$y = x_{11}x_{23} - x_{13}x_{21} \implies \breve{y} = \begin{pmatrix} x_{23} & 0 & -x_{21} \\ -x_{13} & 0 & x_{11} \end{pmatrix},$$

where $\breve{y}$ (produced by the command `\breve{y}`) has the same dimension as X. Generalizing, we have, for any scalar function y of an $n \times q$ matrix X,

$$\breve{y}(X) = \sum_{i=1}^{n} \sum_{j=1}^{q} \frac{\partial y}{\partial x_{ij}} E_{ij}. \tag{3}$$

The matrices $\dot{Y}$ and $\breve{y}$ thus defined contain all the partial derivatives, *but they are not derivatives*. The derivatives of Y and y are

$$DY(x) = \frac{d \operatorname{vec} Y}{dx} = \operatorname{vec} \dot{Y} \tag{4}$$

and

$$Dy(X) = \frac{\partial y}{\partial (\operatorname{vec} X)'} = (\operatorname{vec} \breve{y})', \tag{5}$$

respectively. Dwyer and Macphail (1948) study $\dot{Y}$ and $\breve{y}$, which they call 'symbolic' matrix derivatives, perhaps to emphasize that these are not real

derivatives but just manipulations, and provide some rules for operations with these symbolic derivatives.

As an example, they find the partial derivatives of $Y = AXBX'C$, that is, they work out $\dot{Y}(x_{ij})$ and $\breve{y}_{st}(X)$, which is far from easy using their algebra. In modern notation we simply take the differential

$$dY = A(dX)BX'C + AXB(dX)'C,$$

and then vectorize

$$\begin{aligned} d\,\mathrm{vec}\,Y &= (C'XB' \otimes A)d\,\mathrm{vec}\,X + (C' \otimes AXB)d\,\mathrm{vec}\,X' \\ &= (C'XB' \otimes A + (C' \otimes AXB)K)\,d\,\mathrm{vec}\,X, \end{aligned}$$

using the commutation matrix K (of appropriate order). The derivative is then given by

$$DY(X) = \frac{\partial\,\mathrm{vec}\,Y}{\partial(\mathrm{vec}\,X)'} = C'XB' \otimes A + (C' \otimes AXB)K.$$

Although the work of Dwyer and Macphail was innovative and important, it suffered from two weaknesses. Not only are their 'derivatives' not derivatives, but also their notation does not distinguish between their two definitions. They use $\partial Y/\partial X$ for both cases, specializing to $\partial Y/\partial x_{ij}$ (what we call $\dot{Y}(x_{ij})$) and $\partial y_{st}/\partial X$ (what we call $\breve{y}_{st}(X)$), respectively. Unfortunately, many authors followed their notation, which added to the confusion. The approach of Dwyer and Macphail was practical, not mathematical. Their purpose was to establish a system of principles and rules, which would assist in the practical task of obtaining and organizing the partial derivatives. They were not interested in the underlying mathematical theory, and made no reference to multivariable (vector) calculus or the theory of multilinear algebra. This was unfortunate and has caused much misery.

Almost twenty years passed without much happening.[2] Then, Paul Dwyer extended and simplified the results in Dwyer and Macphail (1948), based in part on results obtained in William Wrobleski's PhD thesis, defended in 1963 under Dwyer's supervision. Dwyer (1967) considers again only scalar functions of a matrix and matrix functions of a scalar, obtains some form of a chain rule, and applies his results to Jacobian determinants, some of them quite intricate (even involving symmetric matrices, where we would

[2]Some papers and books on matrix calculus, including my own *Matrix Differential Calculus* with Neudecker, mention Bodewig's (1959) book, entitled *Matrix Calculus*. This, however, is a book about computational aspects of matrix *algebra*. Matrix *calculus* is not discussed apart from one comment in the (beautifully written) Preface.

now use the duplication matrix). No attempt is made to define a general matrix derivative or to connect the results to the established literature on vector calculus.

An important and crucial step forward was the idea to work via differentials and only at the last step move from differentials to derivatives. This idea was introduced by Heinz Neudecker in 1967 and applied to the 'derivative' $\breve{\phi}$, even managing a preliminary version of the first and second identification theorems; see *Matrix Differential Calculus*, Theorems 5.11 and 6.6. No full treatment is provided in this first attempt, since only scalar functions of a matrix (trace, determinant) are considered.

3 Why bother?

Before we move to the general matrix derivative, why bother? Why is the ordering of the partial derivatives so important? Surely it does not matter how we order the partial derivatives as long as we have all of them collected in one matrix. But it *does* matter, as I tried to explain in Magnus (2010). The derivative is not just a collection of partial derivatives, it is a mathematical object with a meaning, just like the inverse. Consider the following three matrices:

$$A = \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 3 & -5 \\ -1 & 2 \end{pmatrix}, \qquad C = \begin{pmatrix} -1 & 2 \\ -5 & 3 \end{pmatrix}.$$

Then, $B = A^{-1}$, but the matrix C contains the same elements as B and has the additional property that the components of $\operatorname{vec} C$ are in increasing order. So, why not use C as the inverse of A? This is clearly preposterous, because the inverse is more than a collection of its elements — it is a mathematical entity with a meaning.

Exactly the same is true for the derivative. The derivative is not just a collection of partial derivatives. In particular, we want to be able to use a chain rule, we want to interpret the rank of the derivative, and we want to use its determinant in transformation theorems. This is only possible with a correct definition of matrix derivative.

4 Matrix derivatives

Until the late 1960s nobody had succeeded to develop a viable and mathematically correct calculus for matrices. The derivative of a vector is not a vector but a matrix. But what is the derivative of a matrix? A supermatrix?

But we have no algebra for supermatrices.[3] Intuitively, we wish to keep the matrix structure of Y and X intact when we define the derivative. And this misguided intuition is precisely the reason why it has taken so long to arrive at the correct definition and why there is still so much resistance. Because we can't. We have to impose a one-dimensional ordering on the matrices Y and X, and this means we have to map these matrices into vectors, usually $\operatorname{vec} Y$ and $\operatorname{vec} X$. Since the two matrices are now transformed into vectors, the step from vector calculus to matrix calculus is easy.

Two papers, appearing almost simultaneously in the *Journal of the American Statistical Association* 1969, define three matrix 'derivatives' — matrices containing all partial derivatives of a matrix function Y with respect to an argument matrix X. Heinz Neudecker's paper was the first to appear, shortly followed by the paper of Derrick Tracy and Paul Dwyer. These three matrices are defined as follows.

First, based on the expression $\dot{Y}(x_{ij})$ in (2), we define $\dot{Y}(X)$ by assembling the matrices $\dot{Y}(x_{ij})$ in a block structure,

$$\dot{Y}(X) = \sum_{i=1}^{n}\sum_{j=1}^{q} E_{ij} \otimes \dot{Y}(x_{ij}) = \sum_{ij}\sum_{st} \frac{\partial y_{st}}{\partial x_{ij}}(E_{ij} \otimes E_{st}), \tag{6}$$

which is a partitioned matrix of order $mn \times pq$ where each block is of order $m \times p$ (like Y). Second, based on $\breve{y}_{st}(X)$ in (3), we define $\breve{Y}(X)$ as

$$\breve{Y}(X) = \sum_{s=1}^{m}\sum_{t=1}^{p} E_{st} \otimes \breve{y}_{st}(X) = \sum_{st}\sum_{ij} \frac{\partial y_{st}}{\partial x_{ij}}(E_{st} \otimes E_{ij}), \tag{7}$$

also of order $mn \times pq$, where each block is of order $n \times q$ (like X). Given the properties of the commutation matrix, we have

$$K_{mn}\dot{Y}K_{qp} = \breve{Y}, \tag{8}$$

providing a one-to-one correspondence between the two matrices. For example, if $Y = AXB$, then

$$\dot{Y} = (\operatorname{vec} A)(\operatorname{vec} B')', \qquad \breve{Y} = (\operatorname{vec} A')(\operatorname{vec} B)',$$

neither of which is the derivative of Y. The derivative is $DY(X) = B' \otimes A$,

[3]There exists, however, an algebra for tensor products of which the Kronecker product is a special case; see Pollock (1985), who defines tensors in the context of multilinear algebra.

and in general

$$DY(X) = \frac{\partial \operatorname{vec} Y}{\partial (\operatorname{vec} X)'} = \sum_{ij} \sum_{st} \frac{\partial y_{st}}{\partial x_{ij}} (\operatorname{vec} E_{st})(\operatorname{vec} E_{ij})'$$
$$= \sum_{tj} \sum_{si} \frac{\partial y_{st}}{\partial x_{ij}} (E_{tj} \otimes E_{si}), \tag{9}$$

which is of order $mp \times nq$, where each block is of order $m \times n$ and contains the partial derivative of one column of Y with respect to one column of X.

Neither Neudecker nor Tracy and Dwyer seem to have realized that there is an essential difference between $\dot{Y}$ and $\breve{Y}$ on the one hand, and DY on the other hand. Both papers recognize that the algebra through DY is easier, but they continue to think that the ordering of partial derivatives is up to the author and does not really matter. This view persisted until at least 1985 and is echoed is Daan Nel's 1980 survey paper, where he writes:

> 'Although it does not look so at first glance, the vector rearrangement method [that is, the arrangement through (9)] proved to be a very powerful method for writing the matrix derivative as a simple and recognizable statement in terms of the original matrices.'

A third element of confusion was added by the fact that Neudecker (1969) used the wrong definition for the Jacobian from vector analysis. If y is an $m \times 1$ vector function of an $n \times 1$ vector x, then the derivative $Dy(x)$ is an $m \times n$ matrix, but Neudecker defines it as the transpose, that is as an $n \times m$ matrix, which one might call the 'gradient' matrix. Hence, if $Y = AX$, he finds $DY(X) = I \otimes A'$ rather than $I \otimes A$, and if X happens to be a function of Z, say $X = BZ$, then the chain rule (using the wrong definition) gives $DY(Z) = I \otimes B'A'$ rather than the correct expression $DY(Z) = I \otimes AB$. Neudecker's mistake was copied by Tracy and Dwyer (1969) and, surprisingly, by many papers thereafter.

In the 1970s a large literature followed, mostly concentrating on further properties of $\dot{Y}$ and $\breve{Y}$. Matrix calculus was required in control theory (Vetter, 1970), econometrics (Balestra, 1976), psychology (Bentler and Lee, 1978), statistics (Rogers, 1980; Nel, 1980), and many other fields, but a mathematically correct and easy-to-apply theory of matrix calculus was still lacking. In 1978, Peter Bentler and Sik-Yum Lee wrote:

> 'Although multivariable calculus has become a standard mathematical topic in recent years [...], the differential calculus of matrix functions is not yet a mature enough field to have developed

> general matrix calculus rules that reduce to the standard pragmatic rules of scalar calculus when the matrices involved are of order one. It is by no means an easy task to move from scalar calculus to matrix calculus.'

Their paper is an important step forward, because they use the correct concept of matrix derivative, although they don't use differentials. In 1979, Harold Henderson and Shayle Searle also use the correct concept, and they do use differentials. Unfortunately, both Bentler–Lee and Henderson–Searle confuse gradient and derivative, so that the chain rule

$$DY(Z) = DY(X)\,DX(Z)$$

reads $DY(Z) = DX(Z)\,DY(X)$ in their notation.

The confusion between gradient and derivative, which started with the papers by Neudecker and Tracy–Dwyer in 1969, was finally resolved by Stephen Pollock in his 1979 book where he defines (correctly) the derivative of an $m \times 1$ vector function y with respect to an $n \times 1$ vector x as an $m \times n$ matrix, remarking that 'this definition is at variance with a common convention which arrays the partial derivatives in an $n \times m$ matrix' (p. 75). Pollock also employs the correct generalization from vector to matrix calculus.

Heinz Neudecker and I decided in 1981 to write *Matrix Differential Calculus* and it took us seven years to complete the book. Three years before the book came out, we felt that a statement was required concerning the correct definition of matrix derivative and the beauty and strength of the use of differentials. In our 1985 paper we attempted to convince the reader that there is only one correct definition, and we argued strongly against the use of $\dot{Y}$ and $\breve{Y}$ on the grounds of mathematical correctness and computational efficiency.

Maybe the phrase 'only one correct definition' is too strong. Essential is that we organize both Y and X into vectors, for which we typically employ $\operatorname{vec} Y$ and $\operatorname{vec} X$, the column-by-column vectorization. But we could also use the row-by-row vectorization or indeed any other vectorization. The essence is not what type of vectorization is used, but that we employ a vectorization. The column-by-column vectorization orders the components of X as $(x_{11}, x_{21}, \dots)$, while the row-by-row vectorization (advocated by Pollock, 1985) orders the components of X lexicographically as $(x_{11}, x_{12}, \dots)$, which may be more elegant and logical. If $Y = AXB$ then $\operatorname{vec} Y = (B' \otimes A) \operatorname{vec} X$, and hence

$$\operatorname{vec} Y' = K(B' \otimes A) \operatorname{vec} X = (A \otimes B')K \operatorname{vec} X = (A \otimes B') \operatorname{vec} X'.$$

Realizing that $\operatorname{vec} X'$ is precisely the row-by-row vectorization, the derivative based on the row-by-row vectorization is $A \otimes B'$ rather than $B' \otimes A$, which is perhaps more intuitive. The choice between these alternative orderings of the components is a matter of taste, and $\operatorname{vec} X$ (column-by-column) is now standard. But the fact that the phrase 'only one correct definition' is too strong does not mean that $\dot{Y}$ and $\breve{Y}$ are correct definitions of the derivative. They are not.

The 1985 paper did not end the discussion, reason why I wrote another critical paper twenty-five years later where essentially the same message is delivered (Magnus, 2010). The current note has a different flavor: rather than trying to convince the reader, I have tried to understand and explain the confusion that has clouded the field of matrix calculus for so many decades.

5 Patterned matrices

The correct definition of derivative and the subsequent chain rule make it also easy to deal with patterned matrices. A matrix is 'patterned' or obeys a 'linear structure' (Magnus, 1988) if its elements are subject to a linear restriction, such as symmetry, lower triangularity, or diagonality. In a patterned matrix A there exists a subset of $\operatorname{vec} A$ which contain the essential elements of A. If $\psi(A)$ denotes this subvector of $\operatorname{vec} A$, then there exists a unique matrix Δ such that $\operatorname{vec} A = \Delta\psi(A)$. In the case of symmetry, $\psi(A) = \operatorname{vech}(A)$ and Δ is the duplication matrix D.

From the differential

$$d \operatorname{vec} Y = DY(X)\, d \operatorname{vec} X$$

we would conclude that the derivative is $DY(X)$, but if X turns out to be patterned we immediately have, by the chain rule,

$$d \operatorname{vec} X = \Delta\, d\psi(X),$$

and hence

$$d \operatorname{vec} Y = DY(X)\, \Delta\, d\psi(X),$$

leading to the derivative

$$\frac{\partial \operatorname{vec} Y}{\partial (\psi(X))'} = DY(X)\, \Delta.$$

Acknowledgements

I am grateful to Karim Abadir and Stephen Pollock for discussions on notation, definitions, and many other things that helped clarify my mind.

References

Balestra, P. (1976). *La Dérivation Matricielle*, Collection de l'Institut de Mathématiques Economiques, No. 12. Sirey: Paris.

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

Bodewig, E. (1959). *Matrix Calculus*, 3rd edition. North-Holland: Amsterdam.

Dwyer, P. S. and M. S. Macphail (1948). Symbolic matrix derivatives, *Annals of Mathematical Statistics*, 19, 517–534.

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

Henderson, H. V. and S. R. Searle (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics, *Canadian Journal of Statistics*, 7, 65–81.

Magnus, J. R. (1988). *Linear Structures*. Griffin's Statistical Monographs and Courses, No. 42, Edward Arnold: London and Oxford University Press: New York.

Magnus, J. R. (2010). On the concept of matrix derivative, *Journal of Multivariate Analysis*, 101, 2200–2206.

Magnus, J. R. and H. Neudecker (1985). Matrix differential calculus with applications to simple, Hadamard, and Kronecker products, *Journal of Mathematical Psychology*, 29, 474–492.

Magnus, J. R. and H. Neudecker (1988, 2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, third edition. John Wiley and Sons: Chichester/New York.

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

Neudecker, H. (1967). On matrix procedures for optimizing differentiable scalar functions of matrices, *Statistica Neerlandica*, 21, 101–107.

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

Neudecker, H. (1982). On two germane matrix derivatives, *The Matrix and Tensor Quarterly*, 33, 3–12.

Pollock, D. S. G. (1979). *The Algebra of Econometrics*. John Wiley and Sons: New York.

Pollock, D. S. G. (1985). Tensor products and matrix differential calculus, *Linear Algebra and Its Applications*, 67, 169–193.

Rogers, G. S. (1980). *Matrix Derivatives*. Marcel Dekker: New York.

Tracy, D. S. and P. S. Dwyer (1969). Multivariate maxima and minima with matrix derivatives, *Journal of the American Statistical Association*, 64, 1576–1594.

Turnbull, H. W. (1928). On differentiating a matrix, *Proceedings of the Edinburgh Mathematical Society*, 1(2), 111–128.

Vetter, W. J. (1970). Derivative operations on matrices, *IEEE Transactions on Automatic Control*, 15, 241–244.