



On the concept of matrix derivative[☆]

Jan R. Magnus^{*}

Department of Econometrics & OR, Tilburg University, The Netherlands

ARTICLE INFO

Article history:

Received 24 September 2009

Available online 25 May 2010

AMS subject classifications:

1501

15A57

15A72

Keywords:

Matrix calculus

Matrix derivative

Definitions

ABSTRACT

We discuss how to generalize the concept of vector derivative to matrix derivative, propose two definitions, a ‘broad’ and a ‘narrow’ one, compare the two definitions, and argue in favor of the narrow definition.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

In this note we discuss how to generalize the concept of vector derivative to matrix derivative. There are essentially two ways to do this, and they lead to two definitions: the broad definition (the ω -derivative) and the narrow definition (the α -derivative). Both definitions are used in the literature. We argue in favor of the narrow definition.

2. Vector derivatives

In the one-dimensional case, the equation

$$\lim_{u \rightarrow 0} \frac{f(x+u) - f(x)}{u} = f'(x),$$

defining the derivative of a function f at a point x , is equivalent to

$$f(x+u) = f(x) + f'(x)u + r(u),$$

where the remainder $r(u)$ satisfies $\lim_{u \rightarrow 0} r(u)/u = 0$. Conversely, if there exists a quantity a , depending on x but not on u , such that

$$f(x+u) = f(x) + au + r(u), \quad \lim_{u \rightarrow 0} r(u)/u = 0,$$

then f is differentiable at x with $f'(x) = a$. Differentiability of a function and the possibility to approximate a function by means of an affine function are therefore equivalent properties.

[☆] We thank two referees and an associate editor for their constructive and helpful comments.

^{*} Corresponding address: Department of Econometrics & OR, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands.

E-mail address: magnus@uvt.nl.

Using this fundamental equivalence we may generalize the concept of derivative from scalars to vectors, and this is done in standard mathematics textbooks, usually in a chapter entitled ‘functions of several variables’. More precisely, suppose S is an open subset in \mathbb{R}^n , f maps S into \mathbb{R}^m , and $x \in S$. If there exists a linear transformation A of \mathbb{R}^n into \mathbb{R}^m such that

$$f(x + u) = f(x) + Au + r(u), \quad \lim_{u \rightarrow 0} \frac{|r(u)|}{|u|} = 0,$$

then we say that f is differentiable at x , and we write $Df(x) = A$. Hence, if f is an m -vector function of an n -vector x , then its derivative is an $m \times n$ matrix, also denoted as $\partial f(x)/\partial x'$:

$$Df(x) = \frac{\partial f(x)}{\partial x'} = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \dots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}. \tag{1}$$

The notation $\partial f(x)/\partial x'$ has the advantage of bringing out the dimension: we differentiate m elements of a column with respect to n elements of a row, and the result is an $m \times n$ matrix. But this is just handy notation, it is not conceptual. However, the fact that the partial derivatives $\partial f_s(x)/\partial x_i$ are organized in an $m \times n$ matrix and not, for example, in an $n \times m$ matrix or an mn -vector is conceptual *and it matters*. There is no controversy about this definition. All mathematics texts define vector derivatives in this way.

3. Matrix derivatives: broad definition

Now consider a matrix function F of a matrix of variables X , for example $F(X) = AXB$ or $F(X) = X^{-1}$. How should we define the derivative of F ? Surely this must be a matrix containing all the partial derivatives, and this leads to the ‘broad’ definition of matrix derivative.

Definition 1 (Broad). Let $F = (f_{st})$ be an $m \times p$ matrix function of an $n \times q$ matrix of variables $X = (x_{ij})$. Any $r \times c$ matrix A satisfying $rc = mpnq$ and containing all the partial derivatives $\partial f_{st}(X)/\partial x_{ij}$ is called a derivative of F .

In this definition nothing is said about how the partial derivatives are organized in the matrix A . In order to distinguish between the broad definition and the narrow definition (given in the next section), we shall call a derivative satisfying the broad definition an ω -derivative and we shall use δ instead of ∂ . In some textbooks in statistics and econometrics, in a number of papers, and also at the influential Wikipedia [9] website, the ω -derivative is recommended. For example, if F is a function of a scalar x , then one may define an expression

$$\frac{\delta F(x)}{\delta x} = \begin{pmatrix} \frac{\partial f_{11}(x)}{\partial x} & \dots & \frac{\partial f_{1p}(x)}{\partial x} \\ \vdots & & \vdots \\ \frac{\partial f_{m1}(x)}{\partial x} & \dots & \frac{\partial f_{mp}(x)}{\partial x} \end{pmatrix}, \tag{2}$$

which has the same dimension as F . Based on (2) we may then define the ω -derivative of a matrix with respect to a matrix as

$$\frac{\delta F(X)}{\delta X} = \begin{pmatrix} \frac{\delta F(X)}{\delta x_{11}} & \dots & \frac{\delta F(X)}{\delta x_{1q}} \\ \vdots & & \vdots \\ \frac{\delta F(X)}{\delta x_{n1}} & \dots & \frac{\delta F(X)}{\delta x_{nq}} \end{pmatrix}, \tag{3}$$

or as

$$\frac{\delta F(X)}{\delta X'} = \begin{pmatrix} \frac{\delta F(X)}{\delta x_{11}} & \dots & \frac{\delta F(X)}{\delta x_{n1}} \\ \vdots & & \vdots \\ \frac{\delta F(X)}{\delta x_{1q}} & \dots & \frac{\delta F(X)}{\delta x_{nq}} \end{pmatrix}, \tag{4}$$

or indeed in many other ways.

The ω -derivative in (3) is of order $mn \times pq$, and represents the arrangement described in [3], also discussed by Boik [2] and others. In the special case where $p = q = 1$ we obtain the ω -derivative of a vector f with respect to a vector x , and it is

an $mn \times 1$ column vector instead of an $m \times n$ matrix. Hence this ω -derivative does not have the usual vector derivative (1) as a special case.

The ω -derivative in (4) is of order $mq \times np$, and is used in the Wikipedia [9] lemma on ‘matrix calculus’. Because of the transposition of the indexing, the ω -derivative $\delta f(x)/\delta x'$ of a vector with respect to a vector is equal to the derivative $Df(x) = \partial f(x)/\partial x'$. Hence this ω -derivative *does* have the vector derivative (1) as a special case.

Both ω -derivatives share one rather worrying feature. If we look at the vector derivative (1), then we see that the i th row contains the partial derivatives of the i th function with respect to all variables, and the j th column contains the partial derivatives of all functions with respect to the j th variable. This is the key characteristic of the vector derivative, and it does not carry over to ω -derivatives.

4. Matrix derivatives: narrow definition

If we wish to maintain this key characteristic in generalizing the concept of derivative, then we arrive at the narrow definition.

Definition 2 (Narrow). Let $F = (f_{st})$ be an $m \times p$ matrix function of an $n \times q$ matrix of variables $X = (x_{ij})$. Any $mp \times nq$ matrix A containing all the partial derivatives $\partial f_{st}(X)/\partial x_{ij}$ such that each row contains the partial derivatives of *one* function f_{st} with respect to all variables, and each column contains the partial derivatives of all functions with respect to *one* variable x_{ij} , is called a derivative of F .

Let us call this derivative an α -derivative or simply a derivative. The ordering of the functions and the ordering of the variables can be chosen freely. But any row of the derivative *must* contain the partial derivatives of one function only, and any column *must* contain the partial derivatives with respect to one variable only. Any matrix containing all partial derivatives and satisfying this requirement is an α -derivative; any matrix containing all partial derivatives and not satisfying this requirement is not an α -derivative, but an ω -derivative.

To obtain the α -derivative we need to stack the elements of the function F and also the elements of the argument matrix X . Since the vec -operator is the commonly used stacking operator, we use the vec -operator. There exist other stacking operators (for example, by organizing the elements row-by-row rather than column-by-column), and these could be used equally well as long as this is done consistently. A form of stacking, however, is essential in order to preserve the notion of derivative. We note in passing that all stacking operations are in one-to-one correspondence with each other and connected through permutation matrices. For example, the row-by-row and column-by-column stacking operations are connected through the commutation matrix. There is therefore little advantage in developing the theory of matrix calculus for more than one stacking operation. Thus motivated we define the α -derivative as

$$DF(X) = \frac{\partial \text{vec } F(X)}{\partial (\text{vec } X)'}, \quad (5)$$

which is an $mp \times nq$ matrix. The definition of vector derivative ($p = q = 1$) is a special case of the more general definition of matrix derivative, as of course it should. The definition (5) implies that, if F is a function of a scalar x ($n = q = 1$), then $DF(x) = \partial \text{vec } F(x)/\partial x$, a column mp -vector. Also, if φ is a scalar function of a matrix X ($m = p = 1$), then $D\varphi(X) = \partial \varphi(X)/\partial (\text{vec } X)'$, a $1 \times nq$ row vector. The choice of ordering the partial derivatives is not arbitrary. For example, the derivative of the scalar function $\varphi(X) = \text{tr}(X)$ is not $D\varphi(X) = I_n$ (as is often stated), but $D\varphi(X) = (\text{vec } I_n)'$.

5. Does it matter?

Some authors maintain that whether one uses α -derivatives or ω -derivatives is immaterial and just a matter of preference. And it is true: we cannot *prove* that ω -derivatives are incorrect. On the other hand, should it not be the case that a generalization captures the essence of the concept to be generalized?

If we generalize the concept of the reciprocal $b = 1/a$ from a scalar a to a matrix A , we could define a matrix B whose components are $b_{ij} = 1/a_{ij}$. But somehow this does not capture the essence of reciprocal, which is $ab = ba = 1$. Hence we define B to satisfy $AB = BA = I$. The elements of the matrix $B = A^{-1}$ are not ordered arbitrarily. The inverse is more than a collection of its elements. It is a higher-dimensional unit with a meaning.

If we generalize the univariate normal distribution to a ‘matrix-variate’ normal distribution, how should we proceed? Is it not wise to realize that there already exists a multivariate (vector-variate) normal distribution and that many eminent statisticians have thought about how to generalize the normal distribution from univariate to vector-variate? This is an established part of statistics. Is it not desirable then that the matrix-variate normal distribution is a generalization of the vector-variate normal distribution? For example, would it not be strange if the variance matrix of the matrix-variate distribution would not be symmetric or if each row and each column of the variance matrix would not contain all covariances between *one* variate and all other variates? It is true that we cannot *prove* that another arrangement of the elements in the variance matrix is wrong. All we can do is point out that such an arrangement does not generalize the well-established vector-variate distribution to the matrix-variate distribution, and hence places itself outside mainstream statistics.

The same is true for derivatives. There exists a well-established theory of vector derivatives, and we want to define matrix derivatives within this established theory. The eminent mathematicians who contributed to vector calculus realized not only that the possibility to approximate a given function by means of an affine function is the key property of differentiability, but they also realized that in this framework trivial changes such as relabeling functions or variables has only trivial consequences for the derivative: rows and columns are permuted, but the rank is unchanged and the determinant (in the case of a square derivative) is also unchanged (apart possibly from its sign). This is what the α -derivative achieves and what the ω -derivative does not achieve. The arrangement of the partial derivatives matters because a derivative is more than just a collection of partial derivatives. It is a mathematical concept, a mathematical unit, just like the inverse or the variance of the multivariate normal distribution. The point has been made before; see [6–8,4,5,1], but it seems necessary to make it again.

From a theoretical point of view, ω -derivatives do not generalize the concept of vector derivative to matrix derivative and hence do not allow us to interpret the rank of the derivative or its determinant. But also from a practical point of view there are no advantages in using ω -derivatives. It is true that we can obtain a product rule and a chain rule for the various ω -derivatives, but these are more complicated and less easy to use than the product and chain rules for α -derivatives. It seems then that there are only advantages, and not a single disadvantage, for using the concept of α -derivative, as we illustrate further in the following sections.

6. Two simple matrix functions

Let X be an $n \times q$ matrix. We consider two simple matrix functions: the identity function and the transpose function.

The identity function is given by $F(X) = X$. One would expect that the derivative of this function is the identity matrix, and of course we have $DF(X) = I_{nq}$. But the ω -derivative of the identity function is not the identity matrix. Instead we have

$$\frac{\delta F(X)}{\delta X} = (\text{vec } I_n)(\text{vec } I_q)', \quad \frac{\delta F(X)}{\delta X'} = K_{qn},$$

where K_{qn} denotes the commutation matrix. Note that the first matrix has rank one. The second matrix is not the α -derivative but at least it has full rank.

Next consider the transpose function $F(X) = X'$. The derivative is obtained from

$$d \text{vec } F(X) = d \text{vec } X' = K_{nq} d \text{vec } X,$$

so that $DF(X) = K_{nq}$. We find for our two ω -derivatives:

$$\frac{\delta F(X)}{\delta X} = K_{nq}, \quad \frac{\delta F(X)}{\delta X'} = (\text{vec } I_q)(\text{vec } I_n)'$$

Interestingly, the first ω -derivative does lead to the same result as the α -derivative in this case, because the positions of the partial derivatives $\partial x_{ij} / \partial x_{ij} = 1$ coincide. The positions of $\partial x_{st} / \partial x_{ij}$ do not coincide when $(s, t) \neq (i, j)$, but since these are all zero, this has no effect on the ω -derivative.

7. Product rule

Let $F (m \times p)$ and $G (p \times r)$ be functions of $X (n \times q)$. Then the product rule for differentials is simply

$$d(FG) = (dF)G + F(dG). \tag{6}$$

Applying the vec-operator gives

$$d \text{vec}(FG) = (G' \otimes I_m)d \text{vec } F + (I_r \otimes F)d \text{vec } G,$$

so that

$$\frac{\partial \text{vec}(FG)(X)}{\partial (\text{vec } X)'} = (G' \otimes I_m) \frac{\partial \text{vec } F(X)}{\partial (\text{vec } X)'} + (I_r \otimes F) \frac{\partial \text{vec } G(X)}{\partial (\text{vec } X)'}$$

The product rule for α -derivatives is therefore

$$D(FG)(X) = (G' \otimes I_m)DF(X) + (I_r \otimes F)DG(X). \tag{7}$$

It is possible to obtain a product rule for ω -derivatives. Based on

$$\frac{\delta(FG)(X)}{\delta x_{ij}} = \frac{\delta F(X)}{\delta x_{ij}}G + F \frac{\delta G(X)}{\delta x_{ij}}$$

we find

$$\frac{\delta(FG)(X)}{\delta X} = \frac{\delta F(X)}{\delta X}(I_q \otimes G) + (I_n \otimes F) \frac{\delta G(X)}{\delta X}$$

and

$$\frac{\delta(FG)(X)}{\delta X'} = \frac{\delta F(X)}{\delta X'} (I_n \otimes G) + (I_q \otimes F) \frac{\delta G(X)}{\delta X'}.$$

We note that in the lemma ‘matrix calculus’ of Wikipedia [9], the ω -derivative (4) is defined, but the product rule is stated incorrectly.

8. Chain rule

The chain rule is an essential ingredient without which matrix calculus cannot exist. If $F (m \times p)$ is differentiable at $X (n \times q)$, and $G (l \times r)$ is differentiable at $Y = F(X)$, then the composite function $H(X) = G(F(X))$ is differentiable at X , and

$$DH(X) = (DG(Y))(DF(X)). \tag{8}$$

This is the chain rule for α -derivatives.

By analogy, the chain rule for ω -derivatives starts from

$$\frac{\partial h_{st}(X)}{\partial x_{ij}} = \sum_{\alpha,\beta} \frac{\partial g_{st}(Y)}{\partial y_{\alpha\beta}} \cdot \frac{\partial f_{\alpha\beta}(X)}{\partial x_{ij}},$$

from which we see that

$$\frac{\delta H(X)}{\delta x_{ij}} = \sum_{\alpha,\beta} \frac{\partial f_{\alpha\beta}(X)}{\partial x_{ij}} \cdot \frac{\delta G(Y)}{\delta y_{\alpha\beta}} = \left(\frac{\delta(\text{vec } F(X))'}{\delta x_{ij}} \otimes I_l \right) \frac{\delta G(Y)}{\delta \text{vec } Y}.$$

Hence

$$\frac{\delta H(X)}{\delta X} = \left(\frac{\delta(\text{vec } F(X))'}{\delta X} \otimes I_l \right) \left(I_q \otimes \frac{\delta G(Y)}{\delta \text{vec } Y} \right)$$

and

$$\frac{\delta H(X)}{\delta X'} = \left(\frac{\delta(\text{vec } F(X))'}{\delta X'} \otimes I_l \right) \left(I_n \otimes \frac{\delta G(Y)}{\delta \text{vec } Y'} \right).$$

Again we note that in the lemma ‘matrix calculus’ of Wikipedia [9], based on (4), the chain rule is stated incorrectly.

The two chain rules for ω -derivatives do not look inviting. The chain rule for α -derivatives, on the other hand, is simple and straightforward. For example, suppose we have a scalar function of a square matrix X , $\varphi(X) = \log |X|$. Then,

$$d\varphi(X) = \text{tr } X^{-1} dX = (\text{vec } X'^{-1})' d \text{vec } X,$$

and hence $D\varphi(X) = (\text{vec } X'^{-1})'$. Now we are told that in fact X depends on a vector θ . Then, it follows immediately that

$$D\varphi(\theta) = (\text{vec } X'^{-1})' \frac{\partial \text{vec } X}{\partial \theta'}.$$

The reader may wish to try and apply the chain rule using ω -derivatives. This is not straightforward.

9. Transformations

Finally, let us consider a set of transformations

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 4 & 2 & 2 & 1 \\ 1 & 1 & \gamma & 1 \\ 2 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix},$$

or, for short, $y = Ax$, where γ is left free for the moment. The Jacobian of this transformation is the absolute value of the determinant of the derivative, here $|\partial y / \partial x'| = |A| = 6\gamma - 2$. Now suppose we organize x and y into 2×2 matrices, such that $x = \text{vec } X$ and $y = \text{vec } Y$. So, instead of (x_1, x_2, x_3, x_4) we consider $(x_{11}, x_{21}, x_{12}, x_{22})$, and the same for y . What is now the Jacobian of the transformation from the 2×2 matrix X to the 2×2 matrix $Y = F(X)$? The trivial relabeling should not affect the Jacobian. Since

$$DF(X) = \frac{\partial \text{vec } F(X)}{\partial (\text{vec } X)'} = A,$$

the Jacobian of the transformation is still $|A| = 6\gamma - 2$. (Notice that if we organize the four functions in a different way, for example as $(x_{11}, x_{12}, x_{21}, x_{22})$, then the sign of the determinant may change, but not its absolute value. Hence, relabeling does not affect the Jacobian.)

In contrast, the corresponding ω -derivatives are given by

$$A_1 = \frac{\delta F(X)}{\delta X} = \begin{pmatrix} 4 & 2 & 2 & 2 \\ 1 & 0 & \gamma & 0 \\ 2 & 0 & 1 & 0 \\ 1 & 1 & 1 & 2 \end{pmatrix}, \quad A_2 = \frac{\delta F(X)}{\delta X'} = \begin{pmatrix} 4 & 2 & 2 & 0 \\ 1 & 0 & 1 & 1 \\ 2 & 2 & 1 & 0 \\ \gamma & 0 & 1 & 2 \end{pmatrix},$$

with determinants $|A_1| = 4\gamma - 2$ and $|A_2| = -2\gamma$. The three determinants are, of course, not equal in general. If $\gamma = 1/3$, then the transformation is singular ($|A| = 0$), but $|A_1|$ and $|A_2|$ are both nonzero; if $\gamma = 1/2$, then $|A_1| = 0$ but the transformation is in fact nonsingular; and if $\gamma = 0$, then $|A_2| = 0$ but again the transformation nonsingular. Hence, the ω -derivatives do not provide the correct Jacobian of the transformation.

10. An example

There are many examples that show how easy and straightforward matrix calculus is when the proper tools are applied. Here is one. Consider the scalar function

$$\varphi(X) = \text{tr} AX'BX^{-1},$$

where A and B are matrices of constants and X is a nonsingular $n \times n$ matrix. We want to find the derivative and Hessian matrix of φ . We take differentials

$$\begin{aligned} d\varphi &= \text{tr} A(dX)'BX^{-1} + \text{tr} AX'B(dX^{-1}) \\ &= \text{tr}(X')^{-1}B'(dX)A' - \text{tr} AX'BX^{-1}(dX)X^{-1} \\ &= \text{tr} CdX = (\text{vec } C)'d \text{vec } X, \end{aligned}$$

where

$$C = A'(X')^{-1}B' - X^{-1}AX'BX^{-1}.$$

Hence the derivative is

$$D\varphi(X) = (\text{vec } C)'.$$

To find the Hessian we write

$$\begin{aligned} dC &= A'(d(X')^{-1})B' - (dX^{-1})AX'BX^{-1} - X^{-1}A(dX)'BX^{-1} - X^{-1}AX'BdX^{-1} \\ &= -A'(X')^{-1}(dX)'(X')^{-1}B' + X^{-1}(dX)X^{-1}AX'BX^{-1} - X^{-1}A(dX)'BX^{-1} + X^{-1}AX'BX^{-1}(dX)X^{-1}, \end{aligned}$$

so that

$$\begin{aligned} d^2\varphi &= \text{tr}(dC)(dX) = -\text{tr} A'(X')^{-1}(dX)'(X')^{-1}B'(dX) + \text{tr} X^{-1}(dX)X^{-1}AX'BX^{-1}(dX) \\ &\quad - \text{tr} X^{-1}A(dX)'BX^{-1}(dX) + \text{tr} X^{-1}AX'BX^{-1}(dX)X^{-1}(dX) \\ &= 2 \text{tr} X^{-1}(dX)X^{-1}AX'BX^{-1}(dX) - 2 \text{tr} X^{-1}A(dX)'BX^{-1}(dX) \\ &= 2 \text{tr} P_1(dX)P_2(dX) - 2 \text{tr} Q_1(dX)'Q_2(dX) \\ &= 2(d \text{vec } X)'(Q_1' \otimes Q_2)(d \text{vec } X) - 2(d \text{vec } X)'K_{nn}(P_1' \otimes P_2)(d \text{vec } X), \end{aligned}$$

where

$$P_1 = X^{-1}, \quad P_2 = X^{-1}AX'BX^{-1}, \quad Q_1 = X^{-1}A, \quad Q_2 = BX^{-1}.$$

Since a Hessian matrix must be symmetric, we find

$$H\varphi = Q_1' \otimes Q_2 + Q_1 \otimes Q_2' - K_{nn} (P_1' \otimes P_2 + P_2' \otimes P_1).$$

11. Concluding remarks

In this paper we presented two definitions of matrix derivative: the broad definition (ω -derivative) and the narrow definition (α -derivative). Both are used in the literature. We try and argue in favor of the α -derivative, because in our view it maintains the essence of the derivative concept and also because it is easier to use in practice.

Some authors distinguish between Jacobian matrix and (first) derivative and between Hessian matrix and second derivative. This is a necessary distinction when one uses ω -derivatives, but when one uses α -derivatives then there is no difference between first derivative and Jacobian, and between second derivative and Hessian.

In practice one needs the first derivative of matrix functions F with respect to a matrix argument X , and the second derivative of a scalar function f with respect a matrix argument X . This is because, in practice, second-order derivatives typically appear in optimization problems and these are always univariate.

Very occasionally one might need the Hessian matrix of a matrix function $F = (f_{st})$. There is not really a good notation for this ‘super-Hessian’. Perhaps the best presentation is to simply write the result as $Hf_{st}(X)$. Also very occasionally one might need third- and higher-order derivatives. Here also the expressions become non-transparent. In our view the best way to tackle both problems is to write the expressions in terms of differentials. This always works because the use of differentials does not lead to larger matrices: the matrices F , dF , d^2F , and d^3F all have the same order.

References

- [1] K.M. Abadir, J.R. Magnus, *Matrix Algebra, Econometric Exercises*, vol. 1, Cambridge University Press, New York, 2005.
- [2] R.J. Boik, Newton algorithms for analytic rotation: an implicit function approach, *Psychometrika* 73 (2008) 231–259.
- [3] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester, UK, 1981.
- [4] J.R. Magnus, H. Neudecker, Matrix differential calculus with applications to simple, Hadamard, and Kronecker products, *Journal of Mathematical Psychology* 29 (1985) 474–492.
- [5] J.R. Magnus, H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, second ed., John Wiley and Sons, Chichester, New York, 1999, (paperback).
- [6] Matrix calculus, Wikipedia. http://en.wikipedia.org/wiki/matrix_calculus (last consulted 8 December 2009).
- [7] H. Neudecker, Some theorems on matrix differentiation with special reference to Kronecker matrix products, *Journal of the American Statistical Association* 64 (1969) 953–963.
- [8] D.S.G. Pollock, *The Algebra of Econometrics*, John Wiley & Sons, Chichester, UK, 1979.
- [9] D.S.G. Pollock, Tensor products and matrix differential calculus, *Linear Algebra and its Applications* 67 (1985) 169–193.