Notes on Matrix Calculus

Paul L. Fackler*
North Carolina State University

September 27, 2005

Matrix calculus is concerned with rules for operating on functions of matrices. For example, suppose that an $m \times n$ matrix X is mapped into a $p \times q$ matrix Y. We are interested in obtaining expressions for derivatives such as

$$\frac{\partial Y_{ij}}{\partial X_{kl}},$$

for all i, j and k, l. The main difficulty here is keeping track of where things are put. There is no reason to use subscripts; it is far better instead to use a system for ordering the results using matrix operations.

Matrix calculus makes heavy use of the vec operator and Kronecker products. The vec operator vectorizes a matrix by stacking its columns (it is convention that column rather than row stacking is used). For example, vectorizing the matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

*Paul L. Fackler is an Associate Professor in the Department of Agricultural and Resource Economics at North Carolina State University. These notes are copyrighted material. They may be freely copied for individual use but should be appropriated referenced in published work.

Mail: Department of Agricultural and Resource Economics
NCSU, Box 8109
Raleigh NC, 27695, USA
e-mail: paul_fackler@ncsu.edu
Web-site: http://www4.ncsu.edu/~pfackler/
© 2005, Paul L. Fackler

produces

$$\begin{bmatrix} 1 \\ 3 \\ 5 \\ 2 \\ 4 \\ 6 \end{bmatrix}$$

The Kronecker product of two matrices, A and B, where A is $m \times n$ and B is $p \times q$, is defined as

$$A \otimes B = \begin{bmatrix} A_{11}B & A_{12}B & \dots & A_{1n}B \\ A_{21}B & A_{22}B & \dots & A_{2n}B \\ \dots & \dots & \dots & \dots \\ A_{m1}B & A_{m2}B & \dots & A_{mn}B \end{bmatrix},$$

which is an $mp \times nq$ matrix. There is an important relationship between the Kronecker product and the vec operator:

$$\text{vec}(AXB) = (B^\top \otimes A)\text{vec}(X).$$

This relationship is extremely useful in deriving matrix calculus results.

Another matrix operator that will prove useful is one related to the vec operator. Define the matrix $T_{m,n}$ as the matrix that transforms $\text{vec}(A)$ into $\text{vec}(A^\top)$:

$$T_{m,n}\text{vec}(A) = \text{vec}(A^\top).$$

Note the size of this matrix is $mn \times mn$. $T_{m,n}$ has a number of special properties. The first is clear from its definition; if $T_{m,n}$ is applied to the vec of an $m \times n$ matrix and then $T_{n,m}$ applied to the result, the original vectorized matrix results:

$$T_{n,m}T_{m,n}\text{vec}(A) = \text{vec}(A).$$

Thus

$$T_{n,m}T_{m,n} = \text{I}_{mn}.$$

The fact that

$$T_{n,m} = T_{m,n}^{-1}$$

follows directly. Perhaps less obvious is that

$$T_{m,n} = {T_{n,m}}^\top$$

(also combining these results means that $T_{m,n}$ is an orthogonal matrix).

The matrix operator $T_{m,n}$ is a permutation matrix, i.e., it is composed of 0s and 1s, with a single 1 on each row and column. When premultiplying another matrix, it simply rearranges the ordering of rows of that matrix (postmultiplying by $T_{m,n}$ rearranges columns).

The transpose matrix is also related to the Kronecker product. With A and B defined as above,

$$B \otimes A = T_{p,m}(A \otimes B)T_{n,q}.$$

This can be shown by introducing an arbitrary $n \times q$ matrix C:

$$\begin{aligned} T_{p,m}(A \otimes B)T_{n,q}\text{vec}(C) &= T_{p,m}(A \otimes B)\text{vec}(C^\top) \\ &= T_{p,m}\text{vec}(BC^\top A^\top) \\ &= \text{vec}(ACB^\top) \\ &= (B \otimes A)\text{vec}(C). \end{aligned}$$

This implies that $((B \otimes A) - T_{p,m}(A \otimes B)T_{n,q})\text{vec}(C) = 0$. Because C is arbitrary, the desired result must hold.

An immediate corollary to the above result is that

$$(A \otimes B)T_{n,q} = T_{m,p}(B \otimes A).$$

It is also useful to note that $T_{1,m} = T_{m,1} = I_m$. Thus, if A is $1 \times n$ then $(A \otimes B)T_{n,q} = (B \otimes A)$. When working with derivatives of scalars this can result in considerable simplification.

Turning now to calculus, define the derivative of a function mapping $\Re^n \to \Re^m$ as the $m \times n$ matrix of partial derivatives:

$$[Df]_{ij} = \frac{\partial f_i(x)}{\partial x_j}.$$

For example, the simplest derivative is

$$\frac{dAx}{dx} = A.$$

Using this definition, the usual rules for manipulating derivatives apply naturally if one respects the rules of matrix conformability. The summation rule is obvious:

$$D[\alpha f(x) + \beta g(x)] = \alpha Df(x) + \beta Dg(x),$$

where α and β are scalars. The chain rule involves matrix multiplication, which requires conformability. Given two functions $f : \Re^n \to \Re^m$ and $g : \Re^p \to \Re^n$, the derivative of the composite function is

$$D[f(g(x))] = f'(g(x))g'(x).$$

Notice that this satisfies matrix multiplication conformability, whereas the expression $g'(x)f'(g(x))$ attempts to postmultipy an $n \times p$ matrix by an $m \times n$ matrix. To define a product rule, consider the expression $f(x)^\top g(x)$, where $f, g : \Re^n \to \Re^m$. The derivative is the $1 \times n$ vector given by

$$D[f(x)^\top g(x)] = g(x)^\top f'(x) + f(x)^\top g'(x).$$

Notice that no other way of multiplying g by f' and f by g' would ensure conformability. A more general version of the product rule is defined below.

The product rule leads to a useful result about quadratic functions:

$$\frac{dx^\top Ax}{dx} = x^\top A + x^\top A^\top = x^\top (A + A^\top).$$

When A is symmetric this has the very natural form $dx^\top Ax/dx = 2x^\top A$.

These rules define derivatives for vectors. Defining derivatives of matrices with respect to matrices is accomplished by vectorizing the matrices, so $dA(X)/dX$ is the same thing as $d\text{vec}(A(X))/d\text{vec}(X)$. This is where the the relationship between the vec operator and Kronecker products is useful. Consider differentiating $dx^\top Ax$ with respect to A (rather than with respect to x as above):

$$\frac{d\text{vec}(x^\top Ax)}{d\text{vec}(A)} = \frac{d(x^\top \otimes x^\top)\text{vec}(A)}{d\text{vec}(A)} = (x^\top \otimes x^\top)$$

(the derivative of an $m \times n$ matrix A with respect to itself is I_{mn}).

A more general product rule can be defined. Suppose that $f : \Re^n \to \Re^{m \times p}$ and $g : \Re^n \to \Re^{p \times q}$, so $f(x)g(x) : \Re^n \to \Re^{m \times q}$. Using the relationship between the vec and Kronecker product operators

$$\text{vec}(\mathrm{I}_m f(x)g(x)\mathrm{I}_q) = (g(x)^\top \otimes \mathrm{I}_m)\text{vec}(f(x)) = (\mathrm{I}_q \otimes f(x))\text{vec}(g(x)).$$

A natural product rule is therefore

$$Df(x)g(x) = (g(x)^\top \otimes \mathrm{I}_m)f'(x) + (\mathrm{I}_q \otimes f(x))g'(x).$$

This can be used to determine the derivative of $dA^\top A/dA$ where A is $m \times n$.

$$\text{vec}(A^\top A) = (\mathrm{I}_n \otimes A^\top)\text{vec}(A) = (A^\top \otimes \mathrm{I}_n)\text{vec}(A^\top) = (A^\top \otimes \mathrm{I}_n)T_{m,n}\text{vec}(A).$$

Thus (using the product rule)

$$\frac{dA^\top A}{dA} = (\mathrm{I}_n \otimes A^\top) + (A^\top \otimes \mathrm{I}_n)T_{m,n}.$$

This can be simplified somewhat by noting that

$$(A^\top \otimes \mathrm{I}_n)T_{m,n} = T_{n,n}(\mathrm{I}_n \otimes A^\top).$$

Thus

$$\frac{dA^\top A}{dA} = (\mathrm{I}_{n^2} + T_{n,n})(\mathrm{I}_n \otimes A^\top).$$

The product rule is also useful in determining the derivative of a matrix inverse:

$$\frac{dA^{-1}A}{dA} = (A^\top \otimes \mathrm{I}_n)\frac{dA^{-1}}{dA} + (\mathrm{I}_n \otimes A^{-1}).$$

But $A^{-1}A$ is identically equal to I, so its derivative is identically 0. Thus

$$\frac{dA^{-1}}{dA} = -(A^\top \otimes \mathrm{I}_n)^{-1}(\mathrm{I}_n \otimes A^{-1}) = -(A^{-\top} \otimes \mathrm{I}_n)(\mathrm{I}_n \otimes A^{-1}) = -(A^{-\top} \otimes A^{-1}).$$

It is also useful to have an expression for the derivative of a determinant. Suppose A is $n \times n$ with $|A| \neq 0$. The determinant can be written as the product of the ith row of the adjoint of A (A^*) with the ith column of A:

$$|A| = A^*_{i\cdot}A_{\cdot i}.$$

Recall that the elements of the ith row of A^* are not influenced by the elements in the ith column of A and hence

$$\frac{\partial |A|}{\partial A_{\cdot i}} = A^*_{i\cdot}.$$

To obtain the derivative with respect to all of the elements of A, we can concatenate the partial derivatives with respect to each column of A:

$$\frac{d|A|}{dA} = [A^*_{1\cdot}\ A^*_{2\cdot}\ \ldots\ A^*_{n\cdot}] = |A|\left[[A^{-1}]_{1\cdot}\ [A^{-1}]_{2\cdot}\ \ldots\ [A^{-1}]_{n\cdot}\right] = |A|\mathrm{vec}\left(A^{-\top}\right)^\top.$$

The following result is an immediate consequence

$$\frac{d\ln|A|}{dA} = \mathrm{vec}\left(A^{-\top}\right)^\top.$$

Matrix differentiation results allow us to compute the derivatives of the solutions to certain classes of equilibrium problems. Consider, for example, the solution, x, to a linear complementarity problem LCP(M,q) that solves

$$Mx + q \geq 0, \quad x \geq 0, \quad x^\top (Mx + q) = 0$$

The ith element of x is either exactly equal to 0 or is equal to the ith element of $Mx+q$. Define a diagonal matrix D such that $D_{ii} = 1$ if $x > 0$ and equal 0 otherwise. The solution can then be written as $x = -\hat{M}^{-1}Dq$, where $\hat{M} = DM + I - D$. It follows that

$$\frac{\partial x}{\partial q} = -\hat{M}^{-1}D$$

and that

$$\begin{aligned}
\frac{\partial x}{\partial M} &= \frac{\partial x}{\partial \hat{M}^{-1}} \frac{\partial \hat{M}^{-1}}{\partial \hat{M}} \frac{\partial \hat{M}}{\partial M} \\
&= (-q^\top D \otimes I)(-\hat{M}^{-\top} \otimes \hat{M}^{-1})(I \otimes D) \\
&= q^\top D \hat{M}^{-\top} \otimes \hat{M}^{-1} D \\
&= x^\top \otimes \partial x / \partial q
\end{aligned}$$

Given the prevalence of Kronecker products in matrix derivatives, it would be useful to have rules for computing derivatives of Kronecker products themselves, i.e. $dA \otimes B / dA$ and $dA \otimes B / dB$. Because each element of a Kronecker product involves the product of one element from A multiplied by one element of B, the derivative $dA \otimes B / dA$ must be composed of zeros and the elements of B arranged in a certain fashion. Similarly, the derivative $dA \otimes B / dB$ is composed of zeros and the elements of A arranged in a certain fashion.

It can be verified that $dA \otimes B/dA$ can be written as

$$\frac{dA \otimes B}{dA} = \begin{bmatrix} \Psi_1 & 0 & \dots & 0 \\ \Psi_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \Psi_q & 0 & \dots & 0 \\ 0 & \Psi_1 & \dots & 0 \\ 0 & \Psi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \Psi_q & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Psi_1 \\ 0 & 0 & \dots & \Psi_2 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Psi_q \end{bmatrix} = I_n \otimes \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \dots \\ \Psi_q \end{bmatrix}$$

where

$$\Psi_i = \begin{bmatrix} B_{1i} & 0 & \dots & 0 \\ B_{2i} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ B_{pi} & 0 & \dots & 0 \\ 0 & B_{1i} & \dots & 0 \\ 0 & B_{2i} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & B_{pi} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & B_{1i} \\ 0 & 0 & \dots & B_{2i} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & B_{pi} \end{bmatrix} = I_m \otimes B_{\cdot i}$$

This can be written more compactly as

$$\frac{dA \otimes B}{dA} = (I_n \otimes T_{qm} \otimes I_p)(I_{mn} \otimes vec(B)) = (I_{nq} \otimes T_{mp})(I_n \otimes vec(B) \otimes I_m).$$

Similarly $dA \otimes B/dB$ can be written as

$$\frac{dA \otimes B}{dB} = \begin{bmatrix} \Theta_1 & 0 & \dots & 0 \\ 0 & \Theta_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Theta_1 \\ \Theta_2 & 0 & \dots & 0 \\ 0 & \Theta_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Theta_2 \\ \dots & \dots & \dots & \dots \\ \Theta_q & 0 & \dots & 0 \\ 0 & \Theta_q & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Theta_q \end{bmatrix}$$

where

$$\Theta_i = \begin{bmatrix} A_{i1} & 0 & \dots & 0 \\ 0 & A_{i1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{i1} \\ A_{i2} & 0 & \dots & 0 \\ 0 & A_{i2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{i2} \\ \dots & \dots & \dots & \dots \\ A_{im} & 0 & \dots & 0 \\ 0 & A_{im} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{im} \end{bmatrix}.$$

This can be written more compactly as

$$\frac{dA \otimes B}{dB} = (I_n \otimes T_{qm} \otimes I_p)(vec(A) \otimes I_{pq}) = (T_{pq} \otimes I_{mn})(I_q \otimes vec(A) \otimes I_p).$$

Notice that if either A is a row vector ($m = 1$) or B is a column vector ($q = 1$), the matrix $(I_n \otimes T_{qm} \otimes I_p) = Imnpq$ and hence can be ignored.

To illustrate a use for these relationships, consider the second derivative of $xx^\top$ with respect to x, an n-vector.

$$\frac{dxx^\top}{dx} = x \otimes I_n + I_n \otimes x;$$

hence

$$\frac{d^2xx^\top}{dxdx^\top} = (T_{nn} \otimes I_n)(I_n \otimes vec(I_n) + vec(I_n) \otimes I_n).$$

Another example is

$$\frac{d^2A^{-1}}{dAdA} = (I_n \otimes T_{nn} \otimes I_n)\left[\left(I_{n^2} \otimes vec\left(A^{-1}\right)\right)T_{nn} + \left(vec\left(A^{-\top}\right) \otimes I_{n^2}\right)\right]\left(A^{-\top} \otimes A^{-1}\right)$$

Often, especially in statistical applications, one encounters matrices that are symmetrical. It would not make sense to take a derivative with respect to the i, jth element of a symmetric matrix while holding the j, ith element constant. Generally it is preferable to work with a vectorized version of a symmetric matrix that excludes with the upper or lower portion of the matrix. The vech operator is typically taken to be the column-wise vectorization with the upper portion excluded:

$$\text{vech}(A) = \begin{bmatrix} A_{11} \\ \cdots \\ A_{n,1} \\ A_{22} \\ \cdots \\ A_{n2} \\ \cdots \\ A_{nn-1} \\ A_{nn} \end{bmatrix}.$$

One obtains this by selecting elements of vec(A) and therefore we can write

$$\text{vech}(A) = S_n\text{vec}(A),$$

where S_n is an $n(n+1)/2 \times n^2$ matrix of 0s and 1s, with a single 1 in each row. The vech operator can be applied to lower triangular matrices as well; there is no reason to take derivatives with respect to the upper part of a lower triangular matrix (it can also be applied to the transpose of an upper triangular matrix). The use of the vech operator is also important in efficient computer storage of symmetric and triangular matrices.

To illustrate the use of the vech operator in matrix calculus applications, consider an $n \times n$ symmetric matrix C defined in terms of a lower triangular matrix, L,

$$C = LL^\top.$$

Using the already familiar methods it is easy to see that

$$\frac{dC}{dL} = (\mathrm{I}_n \otimes L)T_{n,n} + (L \otimes \mathrm{I}_n).$$

Using the chain rule

$$\frac{d\mathrm{vech}(C)}{d\mathrm{vech}(L)} = \frac{d\mathrm{vech}(C)}{dC}\frac{dC}{dL}\frac{dL}{d\mathrm{vech}(L)} = S_n \frac{dC}{dL} S_n^\top.$$

Inverting this expression provides an expression for $d\mathrm{vech}(L)/d\mathrm{vech}(C)$.

Related to matrix derivatives is the issue of Taylor expansions of matrix-to-matrix functions. One way to think of matrix derivatives is in terms of multidimensional arrays. An $mn \times pq$ matrix can also be thought of as an $m \times n \times p \times q$ 4-dimensional array. The "reshape" function in MATLAB implements such transformations. The ordering of the individual elements has not change, only the way the elements are indexed.

The dth order Taylor expansion of a function $f(X) : R^{m\times n} \to R^{p\times q}$ at $\tilde{X}$ can be computed in the following way

$f = \mathrm{vec}(f(\tilde{X}))$
$dX = \mathrm{vec}(X - \tilde{X})$
for $i = 1$ to d
{
  $f_i = f^{(i)}(\tilde{X})dX$
  for $j = 2$ to i
    $f_i = \mathrm{reshape}(f_i, mn(pq)^{j-2}, pq)dX/j$
  $f = f + f_i$
}
$f = \mathrm{reshape}(f, m, n)$

The techniques described thus far can be applied to computation of derivatives of common "special" functions. First, consider the derivative of a nonnegative integer power A^i of a square matrix A. Application of the chain rule leads to the recursive definition

$$\frac{dA^i}{dA} = \frac{dA^{i-1}A}{dA} = (A^\top \otimes I)\frac{dA^{i-1}}{dA} + (I \otimes A^{i-1})$$

which can also be expressed as a sum of i terms

$$\frac{dA^i}{dA} = \sum_{j=1}^{i} (A^\top)^{i-j} \otimes A^{j-1}.$$

This result can be used to derive an expression for the derivative of the matrix exponential function, which is defined in the usual way in terms of a Taylor expansion:

$$\exp(A) = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

Thus

$$\frac{d \exp(A)}{dA} = \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{j=1}^{i} (A^\top)^{i-j} \otimes A^{j-1}.$$

The same approach can be applied to the matrix natural logarithm:

$$\ln(A) = -\sum_{i=1}^{\infty} \frac{1}{i} (I - A)^i.$$

Summary of Operator Results

A is $m \times n$, B is $p \times q$, X is defined comformably.

$$A \otimes B = \begin{bmatrix} A_{11}B & A_{12}B & \dots & A_{1n}B \\ A_{21}B & A_{22}B & \dots & A_{2n}B \\ \dots & \dots & \dots & \dots \\ A_{m1}B & A_{m2}B & \dots & A_{mn}B \end{bmatrix}$$

$$(AC \otimes BD) = (A \otimes B)(C \otimes D)$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

$$(A \otimes B)^{\top} = A^{\top} \otimes B^{\top}$$

$$\text{vec}(AXB) = (B^{\top} \otimes A)\text{vec}(X)$$

$$\text{trace}(AX) = \text{vec}(A^{\top})^{\top}\text{vec}(X)$$

$$\text{trace}(AX) = \text{trace}(XA)$$

$$T_{m,n}\text{vec}(A) = \text{vec}(A^{\top})$$

$$T_{n,m}T_{m,n} = \text{I}_{mn}$$

$$T_{n,m} = T_{m,n}^{-1}$$

$$T_{m,n} = T_{n,m}^{\top}$$

$$B \otimes A = T_{p,m}(A \otimes B)T_{n,q}$$

Summary of Differentiation Results

A is $m \times n$, B is $p \times q$, x is $n \times 1$, X is defined comformably.

$$[Df]_{ij} = \frac{df_i(x)}{dx_j}$$

$$\frac{dAx}{dx} = A$$

$$D[\alpha f(x) + \beta g(x)] = \alpha Df(x) + \beta Dg(x)$$

$$D[f(g(x))] = f'(g(x))g'(x)$$

$$D[f(x)g(x)] = (g(x)^\top \otimes \mathrm{I}_m)f'(x) + (\mathrm{I}_p \otimes f(x))g'(x).$$

$$\frac{dx^\top Ax}{dx} = x^\top(A + A^\top)$$

$$\frac{d\mathrm{vec}(x^\top Ax)}{d\mathrm{vec}(A)} = (x^\top \otimes x^\top)$$

$$\frac{dA^\top A}{dA} = (\mathrm{I}_{n^2} + T_{n,n})(\mathrm{I}_n \otimes A^\top)$$

$$\frac{dAA^\top}{dA} = (\mathrm{I}_{m^2} + T_{m,m})(A \otimes \mathrm{I}_m)$$

$$\frac{dx^\top A^\top Ax}{dA} = 2x^\top \otimes x^\top A^\top \qquad \text{(when } x \text{ is a vector)}$$

$$\frac{dx^\top AA^\top x}{dA} = 2x^\top A \otimes x^\top \qquad \text{(when } x \text{ is a vector)}$$

$$\frac{dAXB}{dX} = B^\top \otimes A$$

$$\frac{dA^{-1}}{dA} = -(A^{-\top} \otimes A^{-1})$$

$$\frac{d\ln|A|}{dA} = \mathrm{vec}(A^{-\top})^\top$$

$$\frac{d\mathrm{trace}(AX)}{dX} = \mathrm{vec}(A^\top)^\top$$

$$\frac{dA \otimes B}{dA} = (I_n \otimes T_{qm} \otimes I_p)(I_{mn} \otimes vec(B)) = (I_{nq} \otimes T_{mp})(I_n \otimes vec(B) \otimes I_m).$$

$$\frac{dA \otimes B}{dB} = (I_n \otimes T_{qm} \otimes I_p)(vec(A) \otimes I_{pq}) = (T_{pq} \otimes I_{mn})(I_q \otimes vec(A) \otimes I_p).$$

$$\frac{dxx^\top}{dx} = x \otimes I_n + I_n \otimes x$$

$$\frac{d^2xx^\top}{dxdx^\top} = (T_{nn} \otimes I_n)(I_n \otimes vec(I_n) + vec(I_n) \otimes I_n)$$

$$\frac{dA^i}{dA} = (A^\top \otimes I)\frac{dA^{i-1}}{dA} + (I \otimes A^{i-1}) = \sum_{j=1}^{i} (A^\top)^{i-j} \otimes A^{j-1}.$$

$$\frac{d\exp(A)}{dA} = \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{j=1}^{i} (A^\top)^{i-j} \otimes A^{j-1}.$$