A Different(ial) Way

Matrix Derivatives Again

Steven W. Nydick

University of Minnesota

May 17, 2012

Outline

1. Introduction
   - Notation
   - Matrix Calculus: Idea Two
2. Useful Matrix Algebra Rules
   - Vectorized Operators
   - Patterned Matrices
3. Vector Differential Calculus
   - Continuity and Differentiability
   - Cauchy's Rules and the Hessian
   - Matrix Differential Calculus
4. Basic Differentials and Derivatives
   - Preliminary Results
   - Scalar Functions
   - Vector Functions
   - Matrix Functions
5. Differentials of Special Matrices
   - The Inverse
   - The Exponential and Logarithm
6. References

Notation

- $\mathbf{X}$: A matrix
- $\mathbf{x}$: A vector
- x: A scalar

- $\varphi(x)$, $\varphi(\mathbf{x})$, or $\varphi(\mathbf{X})$: A scalar function
- $\mathbf{f}(x)$, $\mathbf{f}(\mathbf{x})$, or $\mathbf{f}(\mathbf{X})$: A vector function
- $\mathbf{F}(x)$, $\mathbf{F}(\mathbf{x})$, or $\mathbf{F}(\mathbf{X})$: A matrix function

- $\mathbf{x}^T$ or $\mathbf{X}^T$: The transpose of $\mathbf{x}$ or $\mathbf{X}$
- x_{ij}: The element in the ith row and jth column of $\mathbf{X}$
- $(x^T)_{ij}$: The element in the ith row and jth column of $\mathbf{X}^T$

- $\mathrm{D}\,\mathbf{f}(x)$: The derivative of the function $\mathbf{f}(x)$
- $\mathrm{d}\,\mathbf{f}(x)$: The differential of the function $\mathbf{f}(x)$

Basic Idea

Vector calculus is well established, but matrix calculus is difficult.

The paper written by Schöneman took one version of the "Calculus of Vectors" and applied it to matrices:

1. The trace operator was a scalar function (of a matrix), that essentially turned matrices into vectors and computed a dot product between them.
    - $\text{tr}(\mathbf{A}^T\mathbf{X}) = \text{vec}(\mathbf{A})^T \text{vec}(\mathbf{X})$
    - vec is the vectorizing operator, stacking the columns of a matrix to create a very long vector.
2. After applying the trace operator, an important subset of maximization problems could be solved by an application of standard vector calculus rules.

Basic Idea

The current paper is based off of the following idea.

1. First -- the entire treatment used differentials.
   - This would allow a vector function to **remain** a vector instead of turning into a matrix.
2. Second -- the derivative was taken with respect to vec($\mathbf{X}$).
   - This would keep the problem as a vector derivative problem **instead** of a matrix derivative problem.
   - Moreover, by undoing the vec operator, we would retain the correct derivative matrix.

Matrix Algebra in Magnus

There are several matrix algebra properties and matrices that Magnus references through his paper and book.

1. The Kronecker Product
2. The Vec/Vech Operator
3. The Duplication Matrix
4. The Commutation Matrix

I will go through these operators in some depth.

The Kronecker Product

The Kronecker Product: Transforms matrices $\mathbf{A} = m \times n$ and $\mathbf{B} = s \times t$ into a matrix $\mathbf{C} = ms \times nt$.

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{in}\mathbf{B} \\ a_{21}\mathbf{B} & a_{11}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix} \tag{1}$$

The most important Kronecker properties are discussed on pp. 27–28 of Magnus & Neudecker (1999).

The Vec Operator

The Vec Operator: Creates a vector from a matrix by stacking the columns of the matrix.

Assume $\mathbf{A}$ is an $m \times n$ matrix such that:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$$

where $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ are the columns of $\mathbf{A}$. Then:

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} \tag{2}$$

Note that $\text{vec}(\mathbf{A})$ is an $mn \times 1$ column vector.

Vec and Kronecker

The vec operator is related to the Kronecker product as follows.

$$\begin{aligned}\operatorname{vec}(\mathbf{a}\mathbf{b}^T) &= \operatorname{vec}\begin{bmatrix}\mathbf{a}b_1 & \mathbf{a}b_2 & \cdots & \mathbf{a}b_n\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{a}b_1 \\ \mathbf{a}b_2 \\ \vdots \\ \mathbf{a}b_n\end{bmatrix} = \begin{bmatrix}b_1\mathbf{a} \\ b_2\mathbf{a} \\ \vdots \\ b_n\mathbf{a}\end{bmatrix} = \mathbf{b}\otimes\mathbf{a}\end{aligned}$$

Thus, as a basic rule

$$\operatorname{vec}(\mathbf{a}\mathbf{b}^T) = \mathbf{b}\otimes\mathbf{a} \tag{3}$$

where **a** and **b** can be **any size** vectors.

Vec and Kronecker 2

Now, assume that $\mathbf{AXC}$ is a conformable matrix product. Furthermore, let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ be the columns (or rows) of a $q \times q$ identity matrix, where q is the number of columns in $\mathbf{X}$.

Then:

$$\begin{aligned}
\sum_{j=1}^{q}(\mathbf{x}_j \mathbf{e}_j^T) &= \mathbf{x}_1 \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} + \mathbf{x}_2 \begin{bmatrix} 0 & 1 & \cdots & 0 \end{bmatrix} + \cdots + \mathbf{x}_q \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{0} \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_q \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_q \end{bmatrix} \\
&= \mathbf{X}
\end{aligned}$$

A matrix can be written as a sum of a bunch of vectors.

Vec and Kronecker 2

Now, based on the last slide

$$\begin{aligned}\text{vec}(\mathbf{AXC}) &= \text{vec}\left(\mathbf{A}\left(\sum_{j=1}^{q}(\mathbf{x}_j\mathbf{e}_j^T)\right)\mathbf{C}\right) \\ &= \text{vec}\left(\sum_{j=1}^{q}(\mathbf{A}\mathbf{x}_j\mathbf{e}_j^T\mathbf{C})\right) \\ &= \text{vec}\left(\sum_{j=1}^{q}[(\mathbf{A}\mathbf{x}_j)(\mathbf{e}_j^T\mathbf{C})]\right)\end{aligned}$$

because $\mathbf{A}$ and $\mathbf{C}$ are constants, $(\mathbf{A}\mathbf{x}_j)$ is a column vector, and $(\mathbf{e}_j^T\mathbf{C})$ is a row vector.

Vec and Kronecker 2

We can continue deriving:

$$\begin{aligned}
\text{vec}\left(\sum_{j=1}^{q}[(\mathbf{A}\mathbf{x}_j)(\mathbf{e}_j^T\mathbf{C})]\right) &= \sum_{j=1}^{q}\text{vec}[(\mathbf{A}\mathbf{x}_j)(\mathbf{e}_j^T\mathbf{C})] \\
&= \sum_{j=1}^{q}[(\mathbf{e}_j^T\mathbf{C})^T \otimes (\mathbf{A}\mathbf{x}_j)] && \text{by (3)} \\
&= \sum_{j=1}^{q}[(\mathbf{C}^T\mathbf{e}_j) \otimes (\mathbf{A}\mathbf{x}_j)] \\
&= \sum_{j=1}^{q}[(\mathbf{C}^T \otimes \mathbf{A})(\mathbf{e}_j \otimes \mathbf{x}_j)]
\end{aligned}$$

because we can pull a sum outside of the vec operator.

Vec and Kronecker 2

Finally:

$$\begin{aligned}
\sum_{j=1}^{q}[(\mathbf{C}^T \otimes \mathbf{A})(\mathbf{e}_j \otimes \mathbf{x}_j)] &= (\mathbf{C}^T \otimes \mathbf{A}) \sum_{j=1}^{q}(\mathbf{e}_j \otimes \mathbf{x}_j) \\
&= (\mathbf{C}^T \otimes \mathbf{A}) \sum_{j=1}^{q} \text{vec}(\mathbf{x}_j \mathbf{e}_j^T) && \text{by (3)} \\
&= (\mathbf{C}^T \otimes \mathbf{A}) \, \text{vec}\left(\sum_{j=1}(\mathbf{x}_j \mathbf{e}_j^T)\right) \\
&= (\mathbf{C}^T \otimes \mathbf{A}) \, \text{vec}(\mathbf{X}) && (4)
\end{aligned}$$

Therefore, a matrix product can be vectorized such that we only need to perform the vec operator on one matrix.

The Vech Operator

The Vech Operator: Creates a vector from a symmetric matrix by stacking the non-duplicate elements column-wise.

Assume $\mathbf{A}$ is a symmetric, square, $n \times n$ matrix.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{21} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \qquad \text{vech}(\mathbf{A}) = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \\ a_{22} \\ \vdots \\ a_{n2} \\ \vdots \\ a_{nn} \end{bmatrix} \tag{5}$$

The Commutation Matrix

Magnus describes several useful patterned matrices.

One useful matrix: The Commutation Matrix.

$$\mathbf{K}_{mn} \quad \text{such that} \quad \mathbf{K}_{mn}\,\text{vec}(\mathbf{A}_{m\times n}) = \text{vec}(\mathbf{A}^T_{n\times m}) \tag{6}$$

The number of rows and columns of $\mathbf{K}$ correspond to the length of $\text{vec}(\mathbf{A})$, because both $\text{vec}(\mathbf{A})$ and $\text{vec}(\mathbf{A}^T)$ have the same number of elements. Moreover, the unique matrix $\mathbf{K}_{mn}$ (with both mn rows and columns) takes $m \rightarrow n$, or flips the columns to be the rows.

Note: The commutation matrix will **always** be square.

The Commutation Matrix

The commutation matrix changes a mn size vector into a nm size vector, so it is square and of size $mn \times mn$.

Moreover, the commutation matrix is just rearranging the elements of the original vector, so it must be a rearranged identity matrix designed to "pick off" the appropriate elements and put each in the correct place.

For instance:

$$\mathbf{A}_{3\times 2} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \qquad \mathbf{K}_{32} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The Commutation Matrix

Thus:

$$\mathbf{K}_{32}\,\mathrm{vec}(\mathbf{A}) = \mathbf{K}_{32}\,\mathrm{vec}\left(\begin{bmatrix}1 & 2\\ 3 & 4\\ 5 & 6\end{bmatrix}\right) = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}1\\ 3\\ 5\\ 2\\ 4\\ 6\end{bmatrix}$$

$$= \begin{bmatrix}1\\ 2\\ 3\\ 4\\ 5\\ 6\end{bmatrix} = \mathrm{vec}\left(\begin{bmatrix}1 & 3 & 5\\ 2 & 4 & 6\end{bmatrix}\right) = \mathrm{vec}(\mathbf{A}^T)$$

The Commutation Matrix

What is the commutation matrix for arbitrary vec($\mathbf{X}_{m\times n}$)?

Given an $m \times n$ matrix $\mathbf{X}$, vec($\mathbf{X}$) will be $mn \times 1$:

1. Elements $1 - m$ of vec($\mathbf{X}$) will correspond to column 1 of $\mathbf{X}$.
2. Elements $(m+1) - 2m$ of vec($\mathbf{X}$) will correspond to column 2 of $\mathbf{X}$.
3. There will be n of these repeating sequences, one for each column of $\mathbf{X}$.

What are contained in the columns of $\mathbf{K}_{mn}$:

1. The first m columns of $\mathbf{K}_{mn}$ will affect **only** the first m elements of vec($\mathbf{X}$).
2. The second m columns of $\mathbf{K}_{mn}$ will affect **only** the second m elements of vec($\mathbf{X}$).
3. There will be n of these blocks.

The Commutation Matrix

So the commutation matrix contains n column blocks each affecting a particular column of $\mathbf{X}$ and corresponding to a particular set of m elements in $\text{vec}(\mathbf{X})$.

$$\left[\begin{array}{cccc|cccc|c|ccc} \mathbf{k}_1 & \mathbf{k}_2 & \cdots & \mathbf{k}_m & \mathbf{k}_{m+1} & \cdots & \mathbf{k}_{m2} & & \cdots & \mathbf{k}_{m(n-1)+1} & \cdots & \mathbf{k}_{mn} \end{array}\right]$$

The vertical lines separate the elements in different columns of $\mathbf{X}$, and each of the $\mathbf{k}_i$ are elementary vectors. Why?

The Commutation Matrix

Now where does each element of a particular block go in the *new* matrix?

We are turning $\text{vec}(\mathbf{X})$ into $\text{vec}(\mathbf{X}^T)$

1. There are n rows (and m columns) in $\mathbf{X}^T$.
2. The first column block of $\mathbf{K}_{mn}$ takes the first column and puts it in the first row.
3. The second column block of $\mathbf{K}_{mn}$ takes the second column and puts it in the second row.
4. Because there are n rows in $\mathbf{X}^T$, elements in the first column of $\mathbf{X}$ (directly next to each other in $\text{vec}(\mathbf{X})$) are now separated by n elements in $\text{vec}(\mathbf{X}^T)$.
5. Because there are n rows in $\mathbf{X}^T$, elements in the second column of $\mathbf{X}$ (directly next to each other in $\text{vec}(\mathbf{X})$) are now separated by n elements in $\text{vec}(\mathbf{X}^T)$.

The Commutation Matrix

Therefore:

1. The columns of $\mathbf{K}_{mn}$ affect the elements of $\text{vec}(\mathbf{X})$, in order.
2. The rows of $\mathbf{K}_{mn}$ represent the particular place of $\text{vec}(\mathbf{X}^T)$, in order.
3. For the first column block of $\mathbf{K}_{mn}$ (affecting the first column of $\mathbf{X}$), there are n rows separating each element in $\mathbf{X}^T$.

So to create a commutation matrix...

1. Create an $mn \times mn$ size matrix
2. Divide the matrix into blocks of m columns
    - Write a line separating column m from column $m + 1$ and column $2m$ from column $2m + 1$, etc.
    - There will be n such column blocks.

The Commutation Matrix

3. Divide the matrix into blocks of n rows.
   - Write a line separating row n from row $n+1$ and row $2n$ from row $2n+1$, etc.
   - There will be m such row blocks.
4. The first n entries of vec($\mathbf{X}^T$) (corresponding to the first n rows of $\mathbf{K}_{mn}$) will be the elements directly to the right of the column separators.
5. The second n entries of vec($\mathbf{X}^T$) (corresponding to the second n rows of $\mathbf{K}_{mn}$) will be the elements one column to the right of the column separators, etc.

The Commutation Matrix

Or, as an example:

$$
\mathbf{K}_{mn} = \left[\begin{array}{cccc|cccc|c|cccc}
1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0 \\
\hline
0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 1 & \cdots & 0 \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\hline
0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 1
\end{array}\right]
$$

The Commutation Matrix

Now let $\mathbf{B}$ be a $p \times q$ matrix, $\mathbf{X}$ be a $q \times n$ matrix, and $\mathbf{A}$ be a $m \times n$ matrix. Then

$$\begin{aligned}
\mathbf{K}_{pm}\,\mathrm{vec}([\mathbf{BXA}^T]_{p\times m}) &= \mathrm{vec}[(\mathbf{BXA}^T)^T] \\
&= \mathrm{vec}(\mathbf{AX}^T\mathbf{B}^T) \\
&= (\mathbf{B} \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{X}^T) && \text{by (4)} \\
&= (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_{qn}\,\mathrm{vec}(\mathbf{X}_{q\times n}) && \text{by (6)}
\end{aligned}$$

But because

$$\mathbf{K}_{pm}\,\mathrm{vec}(\mathbf{BXA}^T) = \mathbf{K}_{pm}(\mathbf{A} \otimes \mathbf{B})\,\mathrm{vec}(\mathbf{X}) \qquad \text{by (4)}$$

it follows that

$$(\mathbf{B} \otimes \mathbf{A})\mathbf{K}_{qn} = \mathbf{K}_{pm}(\mathbf{A} \otimes \mathbf{B}) \tag{7}$$

The Duplication Matrix

Another useful matrix: The Duplication Matrix.

$$\mathbf{D}_n \quad \text{such that} \quad \mathbf{D}_n \,\text{vech}(\mathbf{A}_{n\times n}) = \text{vec}(\mathbf{A}_{n\times n}) \tag{8}$$

The number of rows of $\mathbf{D}$ correspond to the length of vec($\mathbf{A}$), and the number of columns of $\mathbf{D}$ correspond to the length of vech($\mathbf{A}$).

Because vech($\mathbf{A}$) will always be shorter than vec($\mathbf{A}$), $\mathbf{D}$ will have at least as many rows as columns.

Furthermore, the columns of $\mathbf{D}$ are linearly independent. Why?

The Duplication Matrix

The length of vec($\mathbf{A}$) is equal to the number of elements in $\mathbf{A}$, and the length of vec($\mathbf{A}$) is equal to the number of elements on the lower triangle of $\mathbf{A}$.

- The number of rows of $\mathbf{D}$ is equal to n^2.
  - n corresponds to the number of rows/columns of $\mathbf{A}$.
- The number of columns of $\mathbf{D}$ is equal to $[n(n+1)/2]$.

Therefore:

$$\text{Rows of } \mathbf{D}_n = n^2 \qquad\qquad \text{Columns of } \mathbf{D}_n = \frac{n(n+1)}{2}$$

The Duplication Matrix

How does the duplication matrix appear?

- Each column corresponding to an "off-diagonal" element of **A** will have two 1s.
- Each column corresponding to a "diagonal" element of **A** will only have one 1.

$$\mathbf{A}_{3\times 3} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix} \qquad \mathbf{D}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The Duplication Matrix

Why? Well, multiplying vech($\mathbf{A}$) by $\mathbf{D}_3$

$$\mathbf{D}_3 \operatorname{vech}(\mathbf{A}) = \mathbf{D}_3 \operatorname{vech}\left(\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 & 3 & 2 & 4 & 5 & 3 & 5 & 6 \end{bmatrix}^T = \operatorname{vec}\left(\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}\right)$$

$$= \operatorname{vec}(\mathbf{A})$$

The Duplication Matrix

What is the duplication matrix for arbitrary vech($\mathbf{X}_{m\times n}$)?

Given an $n \times n$ matrix $\mathbf{X}$, vec($\mathbf{X}$) will be $[n(n+1)/2] \times 1$:

1. The first n elements of vech($\mathbf{X}$) will correspond to the first column of $\mathbf{X}$.
2. The next $n-1$ elements of vech($\mathbf{X}$) will correspond to the second column of $\mathbf{X}$.
3. The next $n-2$ elements of vech($\mathbf{X}$) will correspond to the third column of $\mathbf{X}$.]
4. The last 1 element of vech($\mathbf{X}$) will correspond to the n^{th} column of $\mathbf{X}$.

Note that vech($\mathbf{X}$) will affect ever decreasing elements in the columns.

The Duplication Matrix

So the duplication matrix contains n blocks each affecting a particular column of $\mathbf{X}$ and corresponding to a particular set of elements in vech($\mathbf{X}$).

$$\left[\begin{array}{ccc|ccc|c|c} \mathbf{d}_1 & \cdots & \mathbf{d}_n & \mathbf{d}_{n+1} & \cdots & \mathbf{d}_{n+(n-1)} & \cdots & \mathbf{d}_{[n(n+1)/2]} \end{array}\right]$$

Rather than dividing blocks of the same length, the separators divide blocks of increasingly shortening lengths because the number of elements in vech($\mathbf{X}$) corresponding to a particular column of $\mathbf{X}$ decreases by 1 in each column.

How many elements are in each column of $\mathbf{D}_n$?

The Duplication Matrix

Now where does each element of a particular block go in the *new* matrix?

We are turning vech($\mathbf{X}$) into vec($\mathbf{X}$)

1. There are n rows and n columns of $\mathbf{X}$.
2. The first column block of $\mathbf{D}_n$ takes the first column and puts it in the first column and first row.
3. The second column block of $\mathbf{D}_n$ takes the second column and puts it in the second column and second row.
4. Because there are n rows in $\mathbf{X}$, elements in the first column of $\mathbf{X}$ are now both directly next to each other at one point **and** separated by n elements at another point.

The Duplication Matrix

So to create a duplication matrix...

1. Create an $n^2 \times [n(n+1)/2]$ size matrix.
2. Divide the matrix into column blocks of decreasing size, starting with size n.
    - Write a line separating column n from column $n+1$ and column $n+(n-1)$ from column $n+(n-1)+1$, etc.
    - There will be n such column blocks.

The Duplication Matrix

3. Divide the matrix into row blocks of size n.
    - Write a line separating row n from row $n + 1$ and row $2n$ from row $2n + 1$, etc.
    - There will be n such row blocks.
4. The first n entries of vec($\mathbf{X}$) (corresponding to the first n rows of $\mathbf{D}_n$) will be the first n columns of $\mathbf{D}_n$.
5. The second n entries of vec($\mathbf{X}$) will consist of the second column in the first block of $\mathbf{D}_n$ followed by *all of the* entries in the second block of $\mathbf{D}_n$.
6. The third n entries of vec($\mathbf{X}$) will consist of the third column in the first block of $\mathbf{D}_n$ followed by the second column in the second block of $\mathbf{D}_n$ followed by *all of the* entries in the third block of $\mathbf{D}_n$.

The Duplication Matrix

Or:

$$
\mathbf{D}_n = \left[\begin{array}{cccc|ccc|c|cc|c}
1 & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\
\hline
0 & 1 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & \cdots & 0 & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 & 0 \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\hline
0 & 0 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & \cdots & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 1 & 0 \\
0 & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 1
\end{array}\right]
$$

Patterned Matrix Code

```
commutator <- function(m, n){
  mn <- m*n
  K <- matrix(0, mn, mn)
  index <- 0
  col <- 0
    for(i in 1:n){
      index <- index + 1
      row <- index
      for(j in 1:m){
        col <- col + 1
        K[row, col] <- 1
        row <- row + n
        }
      }
  return(K)
}
```

```
duplicator <- function(n){
  D <- matrix(0, n^2, n*(n + 1)/2)
  index <- n + 1; row <- 0
  for(i in 1:n){
    index <- index - 1; n2 <- n
    col.blocksep <- n - index + 1
    if(index != n){
      for(k in (index + 1):n){
        row <- row + 1
        D[row, col.blocksep] <- 1
        n2 <- n2 - 1
        col.blocksep <- col.blocksep + n2
    }}
    for(j in 1:index){
      row <- row + 1
      col.ident <- col.blocksep + j - 1
      D[row, col.ident] <- 1
  }}
  return(D)
}
```

Continuity and Accumulation

To understand the treatment of matrix calculus, we should review a few definitions.

Continuity:

- $\varphi(c)$ is continuous at c if one of two things hold:
  1. For any $\epsilon > 0$, there exists a $\delta > 0$ so $||u|| < \delta$ forces $||\varphi(c+u) - \varphi(c)|| < \epsilon$
     - Given any small distance past c in the y direction we can find points close to c in the x direction.
     - Only applies to accumulation points.
  2. c is *not* an accumulation point.
- An accumulation point (or a cluster point) is just a limiting point, meaning that $f(c)$ is the limit of the function $f(c+u)$ as $u \rightarrow 0$.
- *Non accumulation points* are also called *isolated points* or *dots in space*, and are automatically continuous, but trivial.

Differentiability and Taylor Series

Taylor Series:

Using the Taylor Series, we can approximate any function with a polynomial of any size.

$$\begin{aligned}\varphi(x) &= \varphi(c) + \frac{\varphi'(c)}{1!}(x-c) + \cdots + \frac{\varphi^{(n)}(c)}{n!}(x-c)^n + \dots \\ &= \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(c)}{k!}(x-c)^k \\ &= \sum_{k=0}^{p} \frac{\varphi^{(k)}(c)}{k!}(x-c)^k + r_c(x-c)\end{aligned}$$

where r_c (the remainder) usually converges at some rate.

Differentiability and Taylor Series

Taylor Series:

Replacing x with $c + u$, so that $u = x - a$ and

$$\begin{aligned}\varphi(c+u) &= \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(c)}{k!}(u) \\ &= \sum_{k=0}^{p} \frac{\varphi^{(k)}(c)}{k!}(u)^k + r_c(u) \\ &= \varphi(c) + u\varphi'(c) + u^2\frac{\varphi''(c)}{2} + r_{2c}(u) \\ &= \varphi(c) + u\varphi'(c) + r_{1c}(u)\end{aligned}$$

The third line: “second-order Taylor formula”
The fourth line: “first-order Taylor formula”

Differentiability and Talor Series

Rewriting the equation on the previous page:

$$\frac{\varphi(c+u)-\varphi(c)}{u} = \varphi'(c) + \frac{r_{1c}(u)}{u}$$

We know, based on calculus that

$$\lim_{u\to 0} \frac{\varphi(c+u)-\varphi(c)}{u} = \varphi'(c)$$

which is the definition of the derivative and implies

$$\lim_{u\to 0} \frac{r_{1c}(u)}{u} = 0$$

Differentiability

Based on two slides ago, we have

$$\varphi(c+u) = \varphi(c) + u\varphi'(c) + r_{1c}(u)$$

so that $\varphi(c) + u\varphi'(c)$ is the best linear approximation to the original function. But the strength of the linear approximation depends on the size of $r_{1c}(u)$.

The first differential:

$$\mathrm{d}\,\varphi(c;u) = u\varphi'(c) \tag{9}$$

Equation (9) is the linear part of $\varphi(c+u) - \varphi(c)$.

Multidimensional Taylor Series

We can expand to linearize a vector function:

$$\begin{aligned}\mathbf{f}(\mathbf{c}+\mathbf{u}) &= \mathbf{f}(\mathbf{c}) + \mathbf{A}(\mathbf{c})\mathbf{u} + r_c(\mathbf{u}) \\ &= \mathbf{f}(\mathbf{c}) + \mathrm{D}\,\mathbf{f}(\mathbf{c})\mathbf{u} + r_c(\mathbf{u})\end{aligned}$$

Now $||\mathbf{u}|| \to 0$, $d\mathbf{f}(\mathbf{c};\mathbf{u}) = \mathrm{D}\,\mathbf{f}(\mathbf{c})\mathbf{u}$ is called the differential, $\mathrm{D}\,\mathbf{f}(\mathbf{c})$ is the first derivative (Jacobian matrix), and $\nabla\mathbf{f}(\mathbf{c}) = \mathrm{D}\,\mathbf{f}(\mathbf{c})^T$ is the Gradient of $\mathbf{f}$ at $\mathbf{c}$.

Letting $||\mathbf{u}|| \to 0$ would be equivalent to setting $\mathbf{w}$ as a unit-length vector, t as a scalar (such that $t\mathbf{w} = \mathbf{u}$) and making $t \to 0$ ... the directional derivative approach.

Properties of the Differential

Note 1: For the differential to make sense, the original function must be defined on a circle $B(\mathbf{c}; r)$ surrounding $\mathbf{c}$ with radius r, and $\mathbf{c} + \mathbf{u} \in B(\mathbf{c}; r)$.

Note 2: If $f : S \rightarrow \mathbb{R}$, where $\mathbf{f}(S)$ is defined for a set S, and $\mathbf{c}$ is an interior point of that set, the function is continuous at $\mathbf{c}$, and each of the partial derivatives exist in some small space surrounding $\mathbf{c}$, then the derivative exists at $\mathbf{c}$.

Note 3: There is only one first derivative, and the rows of the Jacobian are Gradients of a particular partial functions of the vector function $\mathbf{f}$, whereas the columns are the partial derivatives of $\mathbf{f}$ with respect to a particular element of $\mathbf{c}$.

Multivariate Chain Rule

If $\mathbf{h}(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x}))$, $\mathbf{f}(\mathbf{c}) = \mathbf{b}$, and the function $\mathbf{h}$ is differentiable at $\mathbf{c}$

$$\mathrm{D}\,\mathbf{h}(\mathbf{c}) = (\mathrm{D}\,\mathbf{g}(\mathbf{b}))(\mathrm{D}\,\mathbf{f}(\mathbf{c})) \tag{10}$$

Expanding the multivariate chain rule:

$$\begin{pmatrix} \frac{\partial h(\mathbf{c})_1}{\partial c_1} & \dots & \frac{\partial h(\mathbf{c})_1}{\partial c_n} \\ \vdots & \dots & \vdots \\ \frac{\partial h(\mathbf{c})_k}{\partial c_1} & \dots & \frac{\partial h(\mathbf{c})_k}{\partial c_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial g(\mathbf{b})_1}{\partial b_1} & \dots & \frac{\partial g(\mathbf{b})_1}{\partial b_p} \\ \vdots & \dots & \vdots \\ \frac{\partial g(\mathbf{b})_k}{\partial b_1} & \dots & \frac{\partial g(\mathbf{b})_k}{\partial b_p} \end{pmatrix} \begin{pmatrix} \frac{\partial f(\mathbf{c})_1}{\partial c_1} & \dots & \frac{\partial f(\mathbf{c})_1}{\partial c_n} \\ \vdots & \dots & \vdots \\ \frac{\partial f(\mathbf{c})_p}{\partial c_1} & \dots & \frac{\partial f(\mathbf{c})_p}{\partial c_n} \end{pmatrix}$$

Multivariate Chain Rule

Keeping track of the multivariate chain rule is straightforward if remembering that the “partial functions” go down the rows and the “partial values” go across the columns.

If $h = \varphi$, a univariate function, and $\mathbf{f} = \mathbf{f}(t)$, a multivariate function of a scalar, the multivariate chain rule simplifies.

For example:

$$g(\mathbf{x}) = x_1^2 + 2x_2 \qquad\qquad \mathbf{f}(t) = \begin{pmatrix} t + 2\cos(t) \\ \ln(t) \end{pmatrix}$$

Multivariate Chain Rule

Functions:

$$g(\mathbf{x}) = x_1^2 + 2x_2 \qquad \mathbf{f}(t) = \begin{pmatrix} t + 2\cos(t) \\ \ln(t) \end{pmatrix}$$

Method 1:

$$\begin{aligned} \varphi(t) &= g(\mathbf{f}(t)) \\ &= (t + 2\cos(t))^2 + 2(\ln(t)) \\ &= t^2 + 4t\cos(t) + 4\cos^2(t) + 2\ln(t) \end{aligned}$$

So

$$\begin{aligned} \frac{d\varphi(t)}{dt} &= 2t + 4t[-\sin(t)] + 4\cos(t) + 8\cos(t)[-\sin(t)] + 2(1/t) \\ &= 2t - 4t\sin(t) + 4\cos(t) - 8\cos(t)\sin(t) + 2/t \end{aligned}$$

Multivariate Chain Rule

Functions:

$$g(\mathbf{x}) = x_1^2 + 2x_2 \qquad \mathbf{f}(t) = \begin{pmatrix} t + 2\cos(t) \\ \ln(t) \end{pmatrix}$$

Method 2:

$$\varphi(t) = g(\mathbf{f}(t))$$

So

$$\begin{aligned}
\frac{d\varphi(t)}{dt} &= \begin{pmatrix} \frac{\partial g(\mathbf{f}(t))}{\partial f_1(t)} & \frac{\partial g(\mathbf{f}(t))}{\partial f_2(t)} \end{pmatrix} \begin{pmatrix} \frac{\partial f_1(t)}{\partial t} \\ \frac{\partial f_2(t)}{\partial t} \end{pmatrix} \\
&= \begin{pmatrix} 2\big(t + 2\cos(t)\big) & 2 \end{pmatrix} \begin{pmatrix} 1 - 2\sin(t) \\ 1/t \end{pmatrix} \\
&= [2\big(t + 2\cos(t)\big)][1 - 2\sin(t)] + [2][1/t] \\
&= 2t - 4t\sin(t) + 4\cos(t) - 8\cos(t)\sin(t) + 2/t
\end{aligned}$$

Multivariate Chain Rule

By virtue of the multivariate chain rule process

$$\begin{aligned} \mathbf{f} &: \mathbb{R} \longrightarrow \mathbb{R}^m \\ g &: \mathbb{R}^m \longrightarrow \mathbb{R} \\ \varphi &: \mathbb{R} \longrightarrow \mathbb{R} \end{aligned}$$

Therefore, if φ is a scalar function of a scalar but has a vector as an intermediate step, then we have the chain rule from vector calculus.

$$\frac{d\varphi}{dt} = \sum_{i=1}^{m} \left(\frac{\partial g}{\partial x_i} \frac{\partial x_i}{\partial t} \right)$$

Cauchy's Rule of Invariance

Cauchy's Rule of Invariance:

When we apply the chain rule to a composite differential (instead of only a derivative), the distances also sequentially apply.

$$\begin{aligned} \mathrm{d}\, h(c;u) &= \mathrm{D}\, h(c)u && \text{by } (9) \\ &= (\mathrm{D}\, g(b))(\mathrm{D}\, f(c))u && \text{by } (10) \\ &= \mathrm{D}\, g(b)\, \mathrm{d}\, f(c;u) && \text{by } (9) \\ &= \mathrm{d}\, g[b; \mathrm{d}\, f(c;u)] && (11) \end{aligned}$$

Moving a little bit in the u direction moves $f(c)$ up a particular amount, and moving $f(c)$ up a particular amount moves $g(b)$ up a particular amount (because b and hence $g(b)$ depends on $f(c)$).

The Hessian

A real-valued function can be approximated with a 2nd degree polynomial:

$$\varphi(\mathbf{c}+\mathbf{u}) = \varphi(\mathbf{c}) + D(\varphi(\mathbf{c})) + \frac{1}{2}\mathbf{u}^T\mathbf{B}\mathbf{u} + r_{2c}(\mathbf{u}) \tag{12}$$

as long as the remainder converges at a particular rate.

$$\lim_{||\mathbf{u}||\to 0} = \frac{r(\mathbf{u})}{||\mathbf{u}||^2} = 0$$

Properties of the (Second) Differential

Properties of the second differential:

1. The second differential is just the differential of the first differential.
2. The conditions for the second differential (and second derivative) to exist are identical to the conditions for the first differential to exist. We are just pretending that the first differential is our original function.

The Hessian Matrix

Even though only one vector satisfies

$$d\varphi(\mathbf{c};\mathbf{u}) = \mathbf{a}'\mathbf{u}$$

an infinite number of matrices satisfy

$$d^2\varphi(\mathbf{c};\mathbf{u}) = \mathbf{u}^T\mathbf{B}^*\mathbf{u}$$

And the *unique* Hessian is defined as

$$H\varphi(\mathbf{c}) = \frac{1}{2}(\mathbf{B}(\mathbf{c}) + \mathbf{B}(\mathbf{c})^T) \tag{13}$$

Cauchy's Rule of Invariance: Part II

Unfortunately, the second differential is not Cauchy Invariant.

$$\mathrm{d}^2\, h(c; u) \neq \mathrm{d}^2\, g(b; \mathrm{d}\, f(c; u))$$

Why? Well, by the original chain rule, we have

$$h'(c) = g'(f(c)) \cdot f'(c)$$

which implies that $\mathrm{d}\, h(c; u) = \mathrm{d}\, g(b; \mathrm{d}\, f(c; u))$. But when taking the second derivative, the product rule gets in the way:

$$\begin{aligned} h''(c) &= g''(f(c)) \cdot [f'(c)]^2 + g'(f(c)) \cdot f''(c) \\ &\neq g''(f(c)) \cdot [f'(c)]^2 \end{aligned}$$

Cauchy's Rule of Invariance: Part II

In other words:

1. In the original function, u is a constant with respect to c.
2. In the derivative function, $\mathrm{d}\, f(c;u)$ is *no longer* a constant with respect to c.
    - Same reason why we must apply the product rule in the middle of two chain rules.

Therefore

$$\mathrm{d}^2\, h(c;u) = \mathrm{d}^2\, g(b; \mathrm{d}\, f(c;u)) + \mathrm{d}\, g(b; \mathrm{d}^2\, f(c;u)) \tag{14}$$

by applying the product and chain rules to the first differential.

The Transition: Part I

The transition from vector calculus to matrix calculus is straightforward (according to Magnus).

Step 1: First, he addends his notation to consider matrix derivatives:

If for vector derivatives

$$\mathrm{D}\,\mathbf{f}(\mathbf{x}) := \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T}$$

then for matrix derivatives

$$\mathrm{D}\,\mathbf{F}(\mathbf{X}) := \frac{\partial \mathbf{F}(\mathbf{X})}{\partial [\mathrm{vec}(\mathbf{X})]^T}$$

He thus turns a matrix into a vector to apply the vector theory.

The Transition: Part II

Step 2: Second, he addends his vector differentials to apply to matrices.

$$\operatorname{vec}(\mathrm{d}\,\mathbf{F}(\mathbf{C};\mathbf{U})) = \mathrm{d}\operatorname{vec}(\mathbf{F}(\mathbf{C};\mathbf{U})) = \mathbf{A}(\mathbf{C})\operatorname{vec}(\mathbf{U}) \tag{15}$$

Thus, *every* partial derivative is specified, the order is prespecified, and the theory proceeds from the previous section.

The Transition: Part III

Therefore, to find differentials w.r.t. matrices:

1. Apply the vec operator to both sides.
2. Take the differential of both sides.
3. Simplify until $\mathbf{A}(\mathbf{X})\,\mathrm{d}\,\mathrm{vec}(\mathbf{X})$ is on the right side.
   - $\mathbf{A}$ is a Matrix, and $\mathrm{d}\,\mathrm{vec}(\mathbf{X})$ is a vector.
   - $\mathbf{A}$ must **not** depend on $\mathrm{d}\,\mathrm{vec}(\mathbf{X})$.
4. $\mathbf{A}$ is the derivative.
5. Take the differential again.
6. Simplify until $[\mathrm{d}\,\mathrm{vec}(\mathbf{X})]^T\mathbf{B}(\mathbf{X})[\mathrm{d}\,\mathrm{vec}(\mathbf{X})]$
7. $\frac{1}{2}(\mathbf{B}(\mathbf{X}) + \mathbf{B}(\mathbf{X})^T)$ is the Hessian.

Basic Properties of Differentials

The first six differential/derivative rules:

$$\mathrm{d}\,\mathbf{A} = \mathbf{O} \tag{16}$$
$$\mathrm{d}(\alpha\mathbf{F}) = \alpha\,\mathrm{d}\,\mathbf{F} \tag{17}$$
$$\mathrm{d}(\mathbf{F}+\mathbf{G}) = \mathrm{d}\,\mathbf{F} + \mathrm{d}\,\mathbf{G} \tag{18}$$
$$\mathrm{d}\,\mathrm{tr}\,\mathbf{F} = \mathrm{tr}(\mathrm{d}\,\mathbf{F}) \tag{19}$$
$$\mathrm{d}(\mathbf{F}\mathbf{G}) = (\mathrm{d}\,\mathbf{F})\mathbf{G} + \mathbf{F}(\mathrm{d}\,\mathbf{G}) \tag{20}$$
$$\mathrm{d}(\mathbf{F}\otimes\mathbf{G}) = (\mathrm{d}\,\mathbf{F})\otimes\mathbf{G} + \mathbf{F}\otimes(\mathrm{d}\,\mathbf{G}) \tag{21}$$

which are a consequence of the differential being a linear operator on the derivative, and a derivative matrix being a matrix *of* derivatives.

Basic Properties of Differentials

For instance, take Equation (18):

$$\mathrm{d}(\mathbf{F}+\mathbf{G}) = \mathrm{d}\,\mathbf{F} + \mathrm{d}\,\mathbf{G}$$

For an arbitrary element i in the differential vector

$$\mathrm{d}_i(\mathbf{F}+\mathbf{G}) = \mathrm{D}_{i.}(\mathbf{F}+\mathbf{G})^T\mathbf{u}$$

where $\mathrm{D}_{i.}(\mathbf{F}+\mathbf{G})^T$ is the i^{th} row of $\mathrm{D}(\mathbf{F}+\mathbf{G})$. Finally,

$$\begin{aligned}\mathrm{D}_{i.}(\mathbf{F}+\mathbf{G})^T\mathbf{u} &= \sum_j(\mathrm{D}_{ij}(\mathbf{F}+\mathbf{G})u_j)\\ &= \sum_j(\mathrm{D}_{ij}(\mathbf{F})u_j) + \sum_j(\mathrm{D}_{ij}(\mathbf{G})u_j)\\ &= \mathrm{d}_i\,\mathbf{F} + \mathrm{d}_i\,\mathbf{G}\end{aligned}$$

Because linearity applies for an arbitrary element in the differential vector, it holds for the entire vector of differentials.

Basic Properties of Differentials

Now, take Equation (20):

$$\mathrm{d}(\mathbf{F}\mathbf{G}) = (\mathrm{d}\,\mathbf{F})\mathbf{G} + \mathbf{F}(\mathrm{d}\,\mathbf{G})$$

For an arbitrary element i, j

$$\begin{aligned}
(\mathrm{d}(\mathbf{F}\mathbf{G}))_{ij} = \mathrm{d}(\mathbf{F}\mathbf{G})_{ij} &= \mathrm{d}\left(\sum_k f_{ik}g_{kj}\right) \\
&= \sum_k \mathrm{d}(f_{ik}g_{kj}) \\
&= \sum_k [(\mathrm{d}\,f_{ik})g_{kj} + f_{ik}(\mathrm{d}\,g_{kj})] \\
&= \sum_k [(\mathrm{d}\,f_{ik})g_{kj}] + \sum_k [f_{ik}(\mathrm{d}\,g_{kj})] \\
&= [(\mathrm{d}\,\mathbf{F})\mathbf{G}]_{ij} + [\mathbf{F}(\mathrm{d}\,\mathbf{G})]_{ij}
\end{aligned}$$

Therefore, formulas that work on a linear operator of the derivative also work on the differential.

Basic Scalar Functions: $\varphi(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$

Our first function: $\varphi(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$.

Then

$$\begin{aligned} \mathrm{d}\,\varphi(\mathbf{x}) &= \mathrm{d}(\mathbf{a}^T\mathbf{x}) \\ &= \mathbf{a}^T\,\mathrm{d}\,\mathbf{x} && \text{by (17)} \end{aligned}$$

Thus

$$\mathrm{d}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\,\mathrm{d}\,\mathbf{x} \tag{22}$$

$$\mathrm{D}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T \tag{23}$$

Basic Scalar Functions: $\varphi(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$

Our next function: $\varphi(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$.

Then

$$\begin{aligned}
\mathrm{d}\,\varphi(\mathbf{x}) &= \mathrm{d}(\mathbf{x}^T\mathbf{A}\mathbf{x}) \\
&= \mathrm{d}(\mathbf{x}^T)\mathbf{A}\mathbf{x} + \mathbf{x}^T\,\mathrm{d}(\mathbf{A}\mathbf{x}) && \text{by (20)} \\
&= \mathrm{d}(\mathbf{x})^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}\,\mathrm{d}(\mathbf{x}) && \text{by (17)} \\
&= \mathbf{x}^T\mathbf{A}^T\,\mathrm{d}(\mathbf{x}) + \mathbf{x}^T\mathbf{A}\,\mathrm{d}(\mathbf{x}) \\
&= [\mathbf{x}^T(\mathbf{A}^T + \mathbf{A})]\,\mathrm{d}(\mathbf{x})
\end{aligned}$$

Thus

$$\mathrm{d}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = [\mathbf{x}^T(\mathbf{A}^T + \mathbf{A})]\,\mathrm{d}(\mathbf{x}) \tag{24}$$

$$\mathrm{D}(\mathbf{a}^T\mathbf{x}) = \mathbf{x}^T(\mathbf{A}^T + \mathbf{A}) \tag{25}$$

Scalar Functions of Mat 1: $\varphi(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{b}$

Our third function: $\varphi(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{b}$.

Now the differential is with respect to a matrix.

$$\begin{aligned} \mathrm{d}\,\mathrm{vec}[\varphi(\mathbf{X})] &= \mathrm{vec}[\mathrm{d}\,\varphi(\mathbf{X})] && \text{by (15)} \\ &= \mathrm{vec}[\mathrm{d}(\mathbf{a}^T\mathbf{X}\mathbf{b})] \\ &= \mathrm{vec}[\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{b}] && \text{by (17)} \\ &= \mathbf{b}^T \otimes \mathbf{a}^T\,\mathrm{vec}(\mathrm{d}\,\mathbf{X}) && \text{by (4)} \\ &= \mathbf{b}^T \otimes \mathbf{a}^T\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (15)} \end{aligned}$$

Scalar Functions of Mat 1: $\varphi(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{b}$

According to the previous slide, the matrix differential of $\mathbf{a}^T\mathbf{X}\mathbf{b}$:

$$\mathrm{d}\,\mathrm{vec}[\mathbf{a}^T\mathbf{X}\mathbf{b}] = \mathbf{b}^T \otimes \mathbf{a}^T\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \tag{26}$$

$$\mathrm{D}\,\mathrm{vec}[\mathbf{a}^T\mathbf{X}\mathbf{b}] = \mathbf{b}^T \otimes \mathbf{a}^T = (\mathbf{b} \otimes \mathbf{a})^T \tag{27}$$

Notice the (important) use of the vec to Kronecker rule.

Scalar Functions of Mat 2: $\varphi(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a}$

A slightly more complicated function: $\varphi(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a}$

By Equation (15), the matrix differential is as follows.

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}[\varphi(\mathbf{X})] &= \mathrm{vec}[\mathrm{d}\,\varphi(\mathbf{X})] && \text{by (15)} \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a})] \\
&= \mathrm{vec}[\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a} + \mathbf{a}^T\mathbf{X}\,\mathrm{d}(\mathbf{X}^T)\mathbf{a}] && \text{by (20) and (17)} \\
&= \mathrm{vec}[\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a} + \mathbf{a}^T\mathbf{X}(\mathrm{d}\,\mathbf{X})^T\mathbf{a}] \\
&= \mathrm{vec}[\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a} + (\mathbf{a}^T\mathbf{X}(\mathrm{d}\,\mathbf{X})^T\mathbf{a})^T] && \text{Scalar Transpose} \\
&= \mathrm{vec}[\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a} + \mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a}] \\
&= \mathrm{vec}[2\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a}]
\end{aligned}$$

Scalar Functions of Mat 2: $\varphi(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a}$

Finishing:

$$\begin{aligned}
\operatorname{vec}[2\mathbf{a}^T(\mathrm{d}\,\mathbf{X})\mathbf{X}^T\mathbf{a}] &= 2(\mathbf{X}^T\mathbf{a})^T \otimes \mathbf{a}^T \operatorname{vec}(\mathrm{d}\,\mathbf{X}) && \text{by } (4) \\
&= [2(\mathbf{X}^T\mathbf{a})^T \otimes \mathbf{a}^T]\,\mathrm{d}\operatorname{vec}(\mathbf{X}) && \text{by } (15)
\end{aligned}$$

Thus, the matrix differential of $\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a}$ is

$$\begin{aligned}
\mathrm{d}\operatorname{vec}[\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a}] &= 2(\mathbf{X}^T\mathbf{a})^T \otimes \mathbf{a}^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) && (28) \\
\mathrm{D}\operatorname{vec}[\mathbf{a}^T\mathbf{X}\mathbf{b}] &= 2(\mathbf{X}^T\mathbf{a})^T \otimes \mathbf{a}^T = 2(\mathbf{X}^T\mathbf{a} \otimes \mathbf{a})^T && (29)
\end{aligned}$$

Trace Functions

Finding the differential of trace functions use

$$\mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{vec}(\mathbf{A})^T \mathrm{vec}(\mathbf{B}) \tag{30}$$

Why can we use Equation (30)? Well:

$$\begin{aligned} \mathrm{tr}(\mathbf{A}^T\mathbf{B}) &= \sum_{j=1}^{m}\sum_{i=1}^{n}(a_{ij}b_{ij}) \\ &= \mathrm{vec}(\mathbf{A})^T \mathrm{vec}(\mathbf{B}) \end{aligned}$$

Vectorizing a matrix and taking the dot product is equivalently summing the squares of every entry in the matrix.

Trace Functions: $\text{tr}(\mathbf{A}^T\mathbf{X})$

The first trace function: $\text{tr}(\mathbf{A}^T\mathbf{X})$.

$$\begin{aligned} \text{d}[\text{tr}(\mathbf{A}^T\mathbf{X})] &= \text{d}[\text{vec}(\mathbf{A})^T \,\text{vec}(\mathbf{X})] && \text{by (30)} \\ &= \text{vec}(\mathbf{A})^T \,\text{d}\,\text{vec}(\mathbf{X}) && \text{by (17)} \end{aligned}$$

And the matrix differential of $\text{tr}(\mathbf{A}^T\mathbf{X})$:

$$\text{d}[\text{tr}(\mathbf{A}^T\mathbf{X})] = \text{vec}(\mathbf{A})^T \,\text{d}\,\text{vec}(\mathbf{X}) \tag{31}$$

$$\text{D}[\text{tr}(\mathbf{A}^T\mathbf{X})] = \text{vec}(\mathbf{A})^T \tag{32}$$

Trace Functions: $\mathrm{tr}(\mathbf{X}^p)$

The next trace function: $\mathrm{tr}(\mathbf{X}^p)$.

$$\begin{aligned}
\mathrm{d}[\mathrm{tr}(\mathbf{X}^p)] &= \mathrm{tr}[\mathrm{d}(\mathbf{X}^p)] \\
&= \mathrm{tr}[(\mathrm{d}\,\mathbf{X})\mathbf{X}^{p-1} + \mathbf{X}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{p-2} + \cdots + \mathbf{X}^{p-1}(\mathrm{d}\,\mathbf{X})] && \text{by (20)} \\
&= \mathrm{tr}[(\mathrm{d}\,\mathbf{X})\mathbf{X}^{p-1}] + \mathrm{tr}[\mathbf{X}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{p-2}] + \cdots + \mathrm{tr}[\mathbf{X}^{p-1}(\mathrm{d}\,\mathbf{X})] \\
&= \mathrm{tr}[\mathbf{X}^{p-1}(\mathrm{d}\,\mathbf{X})] + \mathrm{tr}[\mathbf{X}^{p-1}(\mathrm{d}\,\mathbf{X})] + \cdots + \mathrm{tr}[\mathbf{X}^{p-1}(\mathrm{d}\,\mathbf{X})] \\
&= p\ \mathrm{tr}[\mathbf{X}^{p-1}(\mathrm{d}\,\mathbf{X})] \\
&= p\ \mathrm{vec}([\mathbf{X}^T]^{p-1})^T\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (30)}
\end{aligned}$$

Both the second to third line and the third to fourth line use typical trace rules (e.g., “linearity of traces” and “cyclic permutation”).

Trace Functions: $\mathrm{tr}(\mathbf{X}^p)$

And the matrix differential of $\mathrm{tr}(\mathbf{X}^p)$:

$$\mathrm{d}[\mathrm{tr}(\mathbf{X}^p)] = p\ \mathrm{vec}([\mathbf{X}^T]^{p-1})^T\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \tag{33}$$

$$\mathrm{D}[\mathrm{tr}(\mathbf{X}^p)] = p\ \mathrm{vec}([\mathbf{X}^T]^{p-1})^T \tag{34}$$

Note that Equation (34) is similar to differentiating a polynomial scalar.

Trace Functions: $\operatorname{tr}(\mathbf{X}^T\mathbf{X})$

The first *power trace* to differentiate: $\operatorname{tr}(\mathbf{X}^T\mathbf{X})$.

$$
\begin{aligned}
\mathrm{d}\operatorname{tr}(\mathbf{X}^T\mathbf{X}) &= \operatorname{tr}[\mathrm{d}(\mathbf{X}^T\mathbf{X})] \\
&= \operatorname{tr}[\mathrm{d}(\mathbf{X}^T)\mathbf{X} + \mathbf{X}^T\,\mathrm{d}(\mathbf{X})] && \text{by (20)} \\
&= \operatorname{tr}[\mathrm{d}(\mathbf{X})^T\mathbf{X}] + \operatorname{tr}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})] \\
&= \operatorname{tr}[(\mathrm{d}(\mathbf{X})^T\mathbf{X})^T] + \operatorname{tr}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})] \\
&= \operatorname{tr}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})] + \operatorname{tr}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})] \\
&= 2\operatorname{tr}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})]
\end{aligned}
$$

Trace Functions: $\operatorname{tr}(\mathbf{X}^T\mathbf{X})$

And we have

$$\begin{aligned} \mathrm{d}\operatorname{tr}(\mathbf{X}^T\mathbf{X}) &= 2\operatorname{tr}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})] \\ &= 2\operatorname{vec}(\mathbf{X})^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) && \text{by (30)} \end{aligned}$$

which implies

$$\mathrm{d}\operatorname{tr}(\mathbf{X}^T\mathbf{X}) = 2\operatorname{vec}(\mathbf{X})^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) \tag{35}$$

$$\mathrm{D}\operatorname{tr}(\mathbf{X}^T\mathbf{X}) = 2\operatorname{vec}(\mathbf{X})^T \tag{36}$$

Trace Functions: $\operatorname{tr}(\mathbf{XAXB})$

The next *power trace* function: $\operatorname{tr}(\mathbf{XAXB})$.

$$\begin{aligned}
\mathrm{d}\operatorname{tr}(\mathbf{XAXB}) &= \operatorname{tr}[\mathrm{d}(\mathbf{XAXB})] \\
&= \operatorname{tr}[\mathrm{d}(\mathbf{X})\mathbf{AXB} + \mathbf{XA}\,\mathrm{d}(\mathbf{X})\mathbf{B}] && \text{by (20)} \\
&= \operatorname{tr}[\mathbf{AXB}\,\mathrm{d}(\mathbf{X}) + \mathbf{BXA}\,\mathrm{d}(\mathbf{X})] \\
&= \operatorname{tr}[(\mathbf{AXB} + \mathbf{BXA})\,\mathrm{d}(\mathbf{X})] \\
&= \operatorname{vec}[(\mathbf{AXB} + \mathbf{BXA})^T]^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) && \text{by (30)}
\end{aligned}$$

And thus

$$\mathrm{d}\operatorname{tr}(\mathbf{XAXB}) = \operatorname{vec}[(\mathbf{AXB} + \mathbf{BXA})^T]^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) \tag{37}$$

$$\mathrm{D}\operatorname{tr}(\mathbf{XAXB}) = \operatorname{vec}[(\mathbf{AXB} + \mathbf{BXA})^T]^T \tag{38}$$

Trace Functions: $\mathrm{tr}[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p]$

The final scalar function: $\mathrm{tr}[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p]$

This final function (in its more general state) captures the remaining differentials on pages 358–359 of Magnus.

$$\mathrm{tr}(\mathbf{X}^T\mathbf{X}) = \sum_i \sum_j x_{ij}^2 \implies \mathbf{A} = \mathbf{I},\ \mathbf{B} = \mathbf{I},\ \&\ p = 1$$

$$\mathrm{tr}[(\mathbf{X}^T\mathbf{X})^p] = \mathrm{tr}[(\mathbf{X}\mathbf{X}^T)^p] \implies \mathbf{A} = \mathbf{I} \quad \& \quad \mathbf{B} = \mathbf{I}$$

Trace Functions: $\mathrm{tr}[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p]$

First, apply the product rule sequentially:

$$\begin{aligned}\mathrm{d}[\mathrm{tr}\{(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p\}] &= \mathrm{tr}\{\mathrm{d}[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p]\} \\ &= \mathrm{tr}\{\mathrm{d}[\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B}](\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1} \\ &\qquad + \cdots + (\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}\,\mathrm{d}[\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B}]\} \\ &= \ p\ \mathrm{tr}\{(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}\,\mathrm{d}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})\}\end{aligned}$$

Next, note that

$$\mathrm{d}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B}) = (\mathrm{d}\,\mathbf{X})^T\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}$$

which is due to the product rule and multiplication by constants.

Trace Functions: $\mathrm{tr}[(\mathbf{X}^T\mathbf{AXB})^p]$

And replacing:

$$\begin{aligned}
\mathrm{d}[\mathrm{tr}\{(\mathbf{X}^T\mathbf{AXB})^p\}] &= p\ \mathrm{tr}\{(\mathbf{X}^T\mathbf{AXB})^{p-1}\,\mathrm{d}(\mathbf{X}^T\mathbf{AXB})\} \\
&= p\ \mathrm{tr}\{(\mathbf{X}^T\mathbf{AXB})^{p-1}[(\mathrm{d}\,\mathbf{X})^T\mathbf{AXB} + \mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}]\} \\
&= p\ \mathrm{tr}\{(\mathbf{X}^T\mathbf{AXB})^{p-1}(\mathrm{d}\,\mathbf{X})^T\mathbf{AXB} \\
&\qquad + (\mathbf{X}^T\mathbf{AXB})^{p-1}\mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}]\} \\
&= p\ \mathrm{tr}\{[(\mathbf{X}^T\mathbf{AXB})^{p-1}(\mathrm{d}\,\mathbf{X})^T\mathbf{AXB}]^T \\
&\qquad + (\mathbf{X}^T\mathbf{AXB})^{p-1}\mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}]\} \\
&= p\ \mathrm{tr}\{\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T(\mathrm{d}\,\mathbf{X})[(\mathbf{X}^T\mathbf{AXB})^{p-1}]^T \\
&\qquad + (\mathbf{X}^T\mathbf{AXB})^{p-1}\mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}]\} \\
&= p\ \mathrm{tr}\{[(\mathbf{X}^T\mathbf{AXB})^{p-1}]^T\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T(\mathrm{d}\,\mathbf{X}) \\
&\qquad + (\mathbf{X}^T\mathbf{AXB})^{p-1}\mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}]\}
\end{aligned}$$

Trace Functions: $\operatorname{tr}[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p]$

We ultimately have

$$\begin{aligned}
\mathrm{d}[\operatorname{tr}\{(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p\}] &= p\ \operatorname{tr}\{[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T(\mathrm{d}\,\mathbf{X}) \\
&\qquad + (\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}\mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})\mathbf{B}]\} \\
&= p\ \operatorname{tr}\{[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T(\mathrm{d}\,\mathbf{X}) \\
&\qquad + \mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}\mathbf{X}^T\mathbf{A}(\mathrm{d}\,\mathbf{X})]\} \\
&= p\ \operatorname{tr}\{[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T \\
&\qquad + \mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}\mathbf{X}^T\mathbf{A}](\mathrm{d}\,\mathbf{X})\} \\
&= p\ \operatorname{vec}\{\big([(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T \\
&\qquad + \mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}\mathbf{X}^T\mathbf{A}\big)^T\}^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) \\
&= p\ \operatorname{vec}\{\mathbf{A}\mathbf{X}[\mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}] \\
&\qquad + \mathbf{A}^T\mathbf{X}[\mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\}^T\,\mathrm{d}\operatorname{vec}(\mathbf{X})
\end{aligned}$$

Trace Functions: $\mathrm{tr}[(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p]$

Which implies

$$\begin{aligned} \mathrm{d}[\mathrm{tr}\{(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p\}] &= p\ \mathrm{vec}\{\mathbf{A}\mathbf{X}[\mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}] \\ &\qquad + \mathbf{A}^T\mathbf{X}[\mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\}^T\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \quad (39) \\ \mathrm{D}[\mathrm{tr}\{(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^p\}] &= p\ \mathrm{vec}\{\mathbf{A}\mathbf{X}[\mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}] \\ &\qquad + \mathbf{A}^T\mathbf{X}[\mathbf{B}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})^{p-1}]^T\}^T \quad (40) \end{aligned}$$

Even though Equations (39) and (40) do not appear interesting, they generalize to *all* matrix differentials on pages 358–359 of Magnus.

For instance:

$$\begin{aligned} \mathrm{D}[\mathrm{tr}(\mathbf{X}^T\mathbf{X})] &= 1\ \mathrm{vec}\{\mathbf{I}\mathbf{X}[\mathbf{I}(\mathbf{X}^T\mathbf{I}\mathbf{X}\mathbf{I})^0] + \mathbf{I}^T\mathbf{X}[\mathbf{I}(\mathbf{X}^T\mathbf{I}\mathbf{X}\mathbf{I})^0]^T\}^T \\ &= \mathrm{vec}[\mathbf{X}\mathbf{I} + \mathbf{X}\mathbf{I}^T]^T \\ &= \mathrm{vec}[\mathbf{X} + \mathbf{X}]^T = 2\,\mathrm{vec}(\mathbf{X})^T \end{aligned}$$

Trace Differentials

The standard process of computing trace differentials:

1. Put differential *inside* trace operator.
2. Usually perform standard product rule or chain rule.
3. Take transposes and rotate to get $\mathrm{d}(\mathbf{X})$ on the outside.
4. Combine terms.
5. Use the "trace to vec" identity.

Vector Functions of Vec 1: $\mathbf{f}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{x}$

The first vector function: $\mathbf{f}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{x}$.

This is the most general function mentioned on pp. 360 in Magnus.

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{x}$$

If $\mathbf{A}$ depends on $\mathbf{x}$, then

$$\begin{aligned}
\mathrm{d}[\mathbf{f}(\mathbf{x})] &= \mathrm{d}[\mathbf{A}(\mathbf{x})\mathbf{x}] \\
&= \mathrm{d}[\mathbf{A}(\mathbf{x})]\mathbf{x} + \mathbf{A}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} && \text{by (20)} \\
&= \mathrm{vec}\{\mathrm{d}[\mathbf{A}(\mathbf{x})]\mathbf{x}\} + \mathbf{A}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\
&= \mathrm{vec}\{\mathbf{I}\,\mathrm{d}[\mathbf{A}(\mathbf{x})]\mathbf{x}\} + \mathbf{A}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\
&= (\mathbf{x}^T \otimes \mathbf{I})\,\mathrm{vec}\{\mathrm{d}[\mathbf{A}(\mathbf{x})]\} + \mathbf{A}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\
&= (\mathbf{x}^T \otimes \mathbf{I})\,\mathrm{D}\,\mathrm{vec}[\mathbf{A}(\mathbf{x})]\,\mathrm{d}\,\mathbf{x} + \mathbf{A}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} && \text{by (9)} \\
&= \big[(\mathbf{x}^T \otimes \mathbf{I})\,\mathrm{D}\,\mathrm{vec}[\mathbf{A}(\mathbf{x})] + \mathbf{A}(\mathbf{x})\big]\,\mathrm{d}\,\mathbf{x} && (41)
\end{aligned}$$

Vector Functions of Vec 2: $\mathbf{f}(\mathbf{x}) = [\mathbf{x}^T\mathbf{x}]\mathbf{a}(\mathbf{x})$

The second vector function: $\mathbf{f}(\mathbf{x}) = [\mathbf{x}^T\mathbf{x}]\mathbf{a}(\mathbf{x})$.

If $\mathbf{a}$ depends on $\mathbf{x}$, then

$$\begin{aligned}
\mathrm{d}\,\mathbf{f}(\mathbf{x}) &= \mathrm{d}\{[\mathbf{x}^T\mathbf{x}]\mathbf{a}(\mathbf{x})\} \\
&= \mathrm{d}[\mathbf{x}^T\mathbf{x}]\mathbf{a}(\mathbf{x}) + \mathbf{x}^T\mathbf{x}\,\mathrm{d}[\mathbf{a}(\mathbf{x})] && \text{by (20)} \\
&= \mathrm{d}[\mathbf{x}^T\mathbf{I}\mathbf{x}]\mathbf{a}(\mathbf{x}) + \mathbf{x}^T\mathbf{x}\,\mathrm{D}\,\mathbf{a}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\
&= [\mathbf{x}^T(\mathbf{I}^T + \mathbf{I})]\,\mathrm{d}(\mathbf{x})\mathbf{a}(\mathbf{x}) + \mathbf{x}^T\mathbf{x}\,\mathrm{D}\,\mathbf{a}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} && \text{by (24)} \\
&= [2\mathbf{x}^T]\,\mathrm{d}(\mathbf{x})\mathbf{a}(\mathbf{x}) + \mathbf{x}^T\mathbf{x}\,\mathrm{D}\,\mathbf{a}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\
&= 2\mathbf{a}(\mathbf{x})\mathbf{x}^T\,\mathrm{d}(\mathbf{x}) + \mathbf{x}^T\mathbf{x}\,\mathrm{D}\,\mathbf{a}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\
&= [2\mathbf{a}(\mathbf{x})\mathbf{x}^T + \mathbf{x}^T\mathbf{x}\,\mathrm{D}\,\mathbf{a}(\mathbf{x})]\,\mathrm{d}\,\mathbf{x} && (42)
\end{aligned}$$

Vector Functions of Mat: $\mathbf{f}(\mathbf{X}) = \mathbf{X}\mathbf{a}$

A vector function of a matrix: $\mathbf{f}(\mathbf{X}) = \mathbf{X}\mathbf{a}$.

The second example of Magnus ($\mathbf{f}(\mathbf{X}) = \mathbf{X}^T$) is redundant.

$$\begin{aligned}
\mathrm{d}[\mathbf{f}(\mathbf{X})] &= \mathrm{d}[\mathbf{X}\mathbf{a}] \\
&= \mathrm{d}[\mathbf{X}]\mathbf{a} && \text{by (17)} \\
&= \operatorname{vec}(\mathrm{d}[\mathbf{X}]\mathbf{a}) \\
&= \operatorname{vec}(\mathbf{I}\,\mathrm{d}[\mathbf{X}]\mathbf{a}) \\
&= (\mathbf{a}^T \otimes \mathbf{I}_n)\operatorname{vec}(\mathrm{d}\,\mathbf{X}) && \text{by (4)} \\
&= (\mathbf{a}^T \otimes \mathbf{I}_n)\,\mathrm{d}\operatorname{vec}(\mathbf{X}) && (43)
\end{aligned}$$

Matrix Functions of Vec: $\mathbf{F}(\mathbf{x}) = \mathbf{x}\mathbf{x}^T$

The first (and easiest) matrix function: $\mathbf{F}(\mathbf{x}) = \mathbf{x}\mathbf{x}^T$.

To differentiate a matrix, first *vectorize* it:

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= \mathrm{d}\,\mathrm{vec}(\mathbf{x}\mathbf{x}^T) \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{x}\mathbf{x}^T)] \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{x})\mathbf{x}^T + \mathbf{x}\,\mathrm{d}(\mathbf{x}^T)] && \text{by (20)} \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{x})\mathbf{x}^T] + \mathrm{vec}[\mathbf{x}\,\mathrm{d}(\mathbf{x}^T)] \\
&= \mathrm{vec}[\mathbf{I}_n\,\mathrm{d}(\mathbf{x})\mathbf{x}^T] + \mathrm{vec}[\mathbf{x}\,\mathrm{d}(\mathbf{x}^T)\mathbf{I}_n] \\
&= (\mathbf{x} \otimes \mathbf{I}_n)\,\mathrm{d}\,\mathrm{vec}(\mathbf{x}) + (\mathbf{I}_n \otimes \mathbf{x})\,\mathrm{d}\,\mathrm{vec}(\mathbf{x}^T) && \text{by (4)} \\
&= (\mathbf{x} \otimes \mathbf{I}_n)\,\mathrm{d}\,\mathrm{vec}(\mathbf{x}) + (\mathbf{I}_n \otimes \mathbf{x})\,\mathrm{d}\,\mathrm{vec}(\mathbf{x}) \\
&= [(\mathbf{x} \otimes \mathbf{I}_n) + (\mathbf{I}_n \otimes \mathbf{x})]\,\mathrm{d}\,\mathrm{vec}(\mathbf{x}) && (44)
\end{aligned}$$

Matrix Functions of Mat 1: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^2$

A matrix power to differentiate: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^2$ ($\mathbf{X}$ is square).

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\mathbf{F}) = \mathrm{d}\,\mathrm{vec}(\mathbf{X}^2) &= \mathrm{vec}[\mathrm{d}(\mathbf{X}\mathbf{X})] \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X})\mathbf{X} + \mathbf{X}\,\mathrm{d}(\mathbf{X})] && \text{by (20)} \\
&= \mathrm{vec}[\mathbf{I}\,\mathrm{d}(\mathbf{X})\mathbf{X}] + \mathrm{vec}[\mathbf{X}\,\mathrm{d}(\mathbf{X})\mathbf{I}] \\
&= (\mathbf{X}^T \otimes \mathbf{I}_n)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_n \otimes \mathbf{X})\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\
& && \text{by (4)} \\
&= [\mathbf{X}^T \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{X}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && (45)
\end{aligned}$$

Make sure to vectorize first, and the differentials are easy.

Matrix Functions of Mat 2: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T$

Another matrix to differentiate: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T$ ($\mathbf{X}$ is of size $m \times n$).

Just remember that there is a commutation matrix to help.

$$\begin{aligned} \mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= \mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) \\ &= \mathrm{d}[\mathbf{K}_{mn}\,\mathrm{vec}(\mathbf{X})] && \text{by (6)} \\ &= [\mathbf{K}_{mn}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && (46) \end{aligned}$$

Apply the commutation matrix prior to differentiating and then realize that the commutation matrix is a constant with respect to $\mathbf{X}$.

Matrix Functions of Mat 3: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T\mathbf{X}$

Another matrix to differentiate: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T\mathbf{X}$ ($\mathbf{X}$ is of size $m \times n$).

Also taking advantage of the commutation matrix.

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= \mathrm{d}\,\mathrm{vec}(\mathbf{X}^T\mathbf{X}) \\
&= \mathrm{vec}\,\mathrm{d}(\mathbf{X}^T\mathbf{X}) \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X}^T)\mathbf{X} + \mathbf{X}^T\,\mathrm{d}(\mathbf{X})] && \text{by (20)} \\
&= \mathrm{vec}[\mathbf{I}_n\,\mathrm{d}(\mathbf{X}^T)\mathbf{X}] + \mathrm{vec}[\mathbf{X}^T\,\mathrm{d}(\mathbf{X})\mathbf{I}_n] \\
&= (\mathbf{X}^T \otimes \mathbf{I}_n)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) + (\mathbf{I}_n \otimes \mathbf{X}^T)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (4)} \\
&= (\mathbf{X}^T \otimes \mathbf{I}_n)\,\mathrm{d}[\mathbf{K}_{mn}\,\mathrm{vec}(\mathbf{X})] + (\mathbf{I}_n \otimes \mathbf{X}^T)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (6)} \\
&= (\mathbf{X}^T \otimes \mathbf{I}_n)\mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_n \otimes \mathbf{X}^T)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\
&= [(\mathbf{X}^T \otimes \mathbf{I}_n)\mathbf{K}_{mn} + (\mathbf{I}_n \otimes \mathbf{X}^T)]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X})
\end{aligned}$$

Matrix Functions of Mat 3: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T\mathbf{X}$

By Equation (7)

$$\mathbf{K}_{pm}(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_{qn}$$

where $\mathbf{A}$ is of size $m \times n$ and $\mathbf{B}$ is of size $p \times q$.

Therefore, because $\mathbf{X}$ is of size $m \times n$, $\mathbf{X}^T$ is of size $n \times m$, and $\mathbf{I}$ is of size $n \times n$, we have

$$\begin{aligned} \mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= [(\mathbf{X}^T \otimes \mathbf{I}_n)\mathbf{K}_{mn} + (\mathbf{I}_n \otimes \mathbf{X}^T)]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\ &= [\mathbf{K}_{nn}(\mathbf{I}_n \otimes \mathbf{X}^T) + \mathbf{I}_{n^2}(\mathbf{I}_n \otimes \mathbf{X}^T)]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (7)} \\ &= [(\mathbf{K}_{nn} + \mathbf{I}_{n^2})(\mathbf{I}_n \otimes \mathbf{X}^T)]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && (47) \end{aligned}$$

Matrix Functions of Mat 4: $\mathbf{F}(\mathbf{X}) = \mathbf{X}\mathbf{A}\mathbf{X}^T$

Another matrix to differentiate: $\mathbf{F}(\mathbf{X}) = \mathbf{X}\mathbf{A}\mathbf{X}^T$.

If $\mathbf{A}$ is symmetric and $\mathbf{X}$ is of size $m \times n$, then

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= \mathrm{d}\,\mathrm{vec}(\mathbf{X}\mathbf{A}\mathbf{X}^T) \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X}\mathbf{A}\mathbf{X}^T)] \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X})\mathbf{A}\mathbf{X}^T + \mathbf{X}\mathbf{A}\,\mathrm{d}(\mathbf{X}^T)] && \text{by (20)} \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X})\mathbf{A}\mathbf{X}^T] + \mathrm{vec}[\mathbf{X}\mathbf{A}\,\mathrm{d}(\mathbf{X}^T)] \\
&= \mathrm{vec}[\mathbf{I}_m\,\mathrm{d}(\mathbf{X})\mathbf{A}\mathbf{X}^T] + \mathrm{vec}[\mathbf{X}\mathbf{A}\,\mathrm{d}(\mathbf{X}^T)\mathbf{I}_m] \\
&= ([\mathbf{A}\mathbf{X}^T]^T \otimes \mathbf{I}_m)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_m \otimes \mathbf{X}\mathbf{A})\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) && \text{by (4)} \\
&= (\mathbf{X}\mathbf{A} \otimes \mathbf{I}_m)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_m \otimes \mathbf{X}\mathbf{A})\,\mathrm{d}[\mathbf{K}_{mn}\,\mathrm{vec}(\mathbf{X})] && \text{by (6)} \\
&= (\mathbf{X}\mathbf{A} \otimes \mathbf{I}_m)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_m \otimes \mathbf{X}\mathbf{A})\mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X})
\end{aligned}$$

Matrix Functions of Mat 4: $\mathbf{F}(\mathbf{X}) = \mathbf{X}\mathbf{A}\mathbf{X}^T$

Continuing:

$$\begin{aligned} \mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= (\mathbf{X}\mathbf{A} \otimes \mathbf{I}_m)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_m \otimes \mathbf{X}\mathbf{A})\mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\ &= \mathbf{I}_{m^2}(\mathbf{X}\mathbf{A} \otimes \mathbf{I}_m)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + \mathbf{K}_{mm}(\mathbf{X}\mathbf{A} \otimes \mathbf{I}_m)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \quad \text{by (7)} \\ &= [(\mathbf{I}_{m^2} + \mathbf{K}_{mm})(\mathbf{X}\mathbf{A} \otimes \mathbf{I}_m)]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \qquad (48) \end{aligned}$$

Make sure to remember the order of the commutation matrices.

$$\mathbf{K}_{mn}\,\mathrm{vec}(\mathbf{A}_{m\times n}) = \mathrm{vec}(\mathbf{A}^T_{n\times m})$$

$$\mathbf{K}_{pm}(\mathbf{A}_{m\times n} \otimes \mathbf{B}_{p\times q}) = (\mathbf{B}_{p\times q} \otimes \mathbf{A}_{m\times n})\mathbf{K}_{qn}$$

Matrix Functions of Mat 5: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T\mathbf{A}\mathbf{X}^T$

A final matrix function $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T\mathbf{A}\mathbf{X}^T$.

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\mathbf{F}) &= \mathrm{d}\,\mathrm{vec}(\mathbf{X}^T\mathbf{A}\mathbf{X}^T) \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X}^T\mathbf{A}\mathbf{X}^T)] \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X}^T)\mathbf{A}\mathbf{X}^T + \mathbf{X}^T\mathbf{A}\,\mathrm{d}(\mathbf{X}^T)] && \text{by (21)} \\
&= \mathrm{vec}[\mathrm{d}(\mathbf{X}^T)\mathbf{A}\mathbf{X}^T] + \mathrm{vec}[\mathbf{X}^T\mathbf{A}\,\mathrm{d}(\mathbf{X}^T)] \\
&= \mathrm{vec}[\mathbf{I}_n\,\mathrm{d}(\mathbf{X}^T)\mathbf{A}\mathbf{X}^T] + \mathrm{vec}[\mathbf{X}^T\mathbf{A}\,\mathrm{d}(\mathbf{X}^T)\mathbf{I}_m] \\
&= (\mathbf{X}\mathbf{A}^T \otimes \mathbf{I}_n)\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) + (\mathbf{I}_m \otimes \mathbf{X}^T\mathbf{A})\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) && \text{by (4)} \\
&= (\mathbf{X}\mathbf{A}^T \otimes \mathbf{I}_n)\,\mathrm{d}[\mathbf{K}_{mn}\,\mathrm{vec}(\mathbf{X})] + (\mathbf{I}_m \otimes \mathbf{X}^T\mathbf{A})\,\mathrm{d}[\mathbf{K}_{mn}\,\mathrm{vec}(\mathbf{X}^T)] \\
& && \text{by (6)} \\
&= (\mathbf{X}\mathbf{A}^T \otimes \mathbf{I}_n)\mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) + (\mathbf{I}_m \otimes \mathbf{X}^T\mathbf{A})\mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\
&= \{[(\mathbf{X}\mathbf{A}^T \otimes \mathbf{I}_n) + (\mathbf{I}_m \otimes \mathbf{X}^T\mathbf{A})]\mathbf{K}_{mn}\}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && (49)
\end{aligned}$$

Differential of The Inverse: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^{-1}$

The differential of the inverse is a "once seen, never forgotten" problem.

Note that *if* **X** is invertible, then

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

A matrix times its inverse is the identiy matrix.

And the differential of the left size is equivalent to that of the right side:

> *The right side is a constant (**I**), and what is the differential of a constant?*

Differential of The Inverse: $\mathbf{F}(\mathbf{X}) = \mathbf{X}^{-1}$

Because

$$\mathrm{d}(\mathbf{I}) = \mathbf{0} \qquad \text{by (16)}$$

We have

$$\begin{aligned} \mathrm{d}(\mathbf{X}^{-1}\mathbf{X}) &= \mathrm{d}(\mathbf{I}) \\ \mathrm{d}(\mathbf{X}^{-1}\mathbf{X}) &= \mathbf{0} && \text{by (16)} \\ \mathrm{d}(\mathbf{X}^{-1})\mathbf{X} + \mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X}) &= \mathbf{0} && \text{by (21)} \\ \mathrm{d}(\mathbf{X}^{-1})\mathbf{X} &= -\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X}) \\ \mathrm{d}(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1} \end{aligned}$$

And

$$\mathrm{d}(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1} \tag{50}$$

Differential of the Inverse

To differentiate a function that involves the inverse of $\mathbf{X}$:

1. Vectorize everything.
2. Perform standard rules:
    - Multiplication rules
    - Chain rules
    - Pulling out constants
3. Isolate $\mathrm{d}(\mathbf{X}^{-1})$.
4. Perform the diff-of-inverse rule.
5. Separate particular linear combinations and use trace rules
6. Use the vec → Kronecker rule or trace → vec rule.
7. Fiddle with transposes and $\mathbf{K}_{mn}$, and recombine terms.

Inverse Example 1: $\varphi(\mathbf{X}) = \operatorname{tr}(\mathbf{A}\mathbf{X}^{-1})$

An example of inverse differentials: $\varphi(\mathbf{X}) = \operatorname{tr}(\mathbf{A}\mathbf{X}^{-1})$.

$$\begin{aligned} \mathrm{d}\,\varphi(\mathbf{X}) &= \mathrm{d}\operatorname{tr}(\mathbf{A}\mathbf{X}^{-1}) \\ &= \operatorname{tr}[\mathrm{d}(\mathbf{A}\mathbf{X}^{-1})] \end{aligned}$$

Perform standard rules (e.g., multiplication by a constant).

$$\begin{aligned} \mathrm{d}\,\varphi(\mathbf{X}) &= \operatorname{tr}[\mathrm{d}(\mathbf{A}\mathbf{X}^{-1})] \\ &= \operatorname{tr}[\mathbf{A}\,\mathrm{d}(\mathbf{X}^{-1})] \end{aligned}$$

Perform the diff-of-inverse rule.

$$\begin{aligned} \mathrm{d}\,\varphi(\mathbf{X}) &= \operatorname{tr}[\mathbf{A}\,\mathrm{d}(\mathbf{X}^{-1})] \\ &= \operatorname{tr}[\mathbf{A}(-1\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1})] = -1\operatorname{tr}(\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1}) \quad \text{by (50)} \end{aligned}$$

Inverse Example 1: $\varphi(\mathbf{X}) = \operatorname{tr}(\mathbf{A}\mathbf{X}^{-1})$

Use trace identities (e.g., transposing and rotating).

$$\begin{aligned} \mathrm{d}\,\varphi(\mathbf{X}) &= -1\operatorname{tr}(\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1}) \\ &= -\operatorname{tr}[\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})] \end{aligned}$$

And finally, perform the trace → vec rule.

$$\begin{aligned} \mathrm{d}\,\varphi(\mathbf{X}) &= -\operatorname{tr}[\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})] \\ &= -\operatorname{vec}[(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1})^T]^T\,\mathrm{d}\operatorname{vec}(\mathbf{X}) \end{aligned} \tag{51}$$

Inverse Example 2: $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

The next example is a strange matrix function.

$$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$\mathbf{M}$ is <u>idempotent</u> (so the square of itself is itself).

$$\begin{aligned}
\mathbf{M}^2 &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= \mathbf{I}_n^2 - \mathbf{I}_n[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] - [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{I}_n \\
&\quad + [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^2 \\
&= \mathbf{I}_n - 2[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] + [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T][\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \\
&= \mathbf{I}_n - 2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{I}_n - 2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{M}
\end{aligned}$$

Inverse Example 2: $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ might look pretty familiar.

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{52}$$

maps $\mathbf{y}$ into the column space defined by the predictors, but

$$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{53}$$

maps $\mathbf{y}$ into the space orthogonal to the predictors (the error space).

$$\mathbf{H}\mathbf{y} = \hat{\mathbf{y}} \qquad\qquad \mathbf{M}\mathbf{y} = \hat{\boldsymbol{\epsilon}}$$

Inverse Example 2: $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

To find the differential, perform standard (e.g., product) rules.

$$\begin{aligned}
\mathrm{d}(\mathbf{M}) &= \mathrm{d}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= -\,\mathrm{d}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= -[\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}\,\mathrm{d}[(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{X}^T \\
&\quad + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)] && \text{by (21)}
\end{aligned}$$

Concentrate on the inverse differential.

$$\begin{aligned}
\mathrm{d}[(\mathbf{X}^T\mathbf{X})^{-1}] &= -(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} && \text{by (50)} \\
&= -(\mathbf{X}^T\mathbf{X})^{-1}[\mathrm{d}(\mathbf{X}^T)\mathbf{X} + \mathbf{X}^T\,\mathrm{d}(\mathbf{X})](\mathbf{X}^T\mathbf{X})^{-1} && \text{by (21)}
\end{aligned}$$

Inverse Example 2: $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

Plug the differential back in the equation.

$$\begin{aligned}
\mathrm{d}(\mathbf{M}) &= -[\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}\,\mathrm{d}[(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{X}^T \\
&\quad + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)] \\
&= -[\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&\quad + \mathbf{X}[-(\mathbf{X}^T\mathbf{X})^{-1}[\mathrm{d}(\mathbf{X}^T)\mathbf{X} + \mathbf{X}^T\,\mathrm{d}(\mathbf{X})](\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{X}^T \\
&\quad + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)] \\
&= -[\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&\quad - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&\quad - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)] \\
&= -[\mathbf{I}_n\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&\quad + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \\
&= -[\mathbf{M}\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{M}]
\end{aligned}$$

Inverse Example 2: $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

And finally vectorizing everything.

$$\begin{aligned}\mathrm{d}\,\mathrm{vec}(\mathbf{M}) &= -\,\mathrm{vec}[\mathbf{M}\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{M}] \\ &= -\,\mathrm{vec}[\mathbf{M}\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] - \mathrm{vec}[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{M}]\end{aligned}$$

which leads to the vec → Kronecker rule.

$$\begin{aligned}\mathrm{d}\,\mathrm{vec}(\mathbf{M}) &= -\,\mathrm{vec}[\mathbf{M}\,\mathrm{d}(\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] - \mathrm{vec}[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathrm{d}(\mathbf{X}^T)\mathbf{M}] \\ &= -[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \otimes \mathbf{M}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\ &\quad - [\mathbf{M}^T \otimes \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T)\end{aligned}$$

Inverse Example 2: $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

And then applying the commutation matrix,

$$\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) = \mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X})$$

on the above function.

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\mathbf{M}) &= -[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \otimes \mathbf{M}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\
&\quad - [\mathbf{M}^T \otimes \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}^T) \\
&= -[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \otimes \mathbf{M}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\
&\quad - [\mathbf{M} \otimes \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{K}_{mn}\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (6)} \\
&= -\mathbf{I}_{m^2}[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \otimes \mathbf{M}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) \\
&\quad - \mathbf{K}_{mm}[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \otimes \mathbf{M}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (7)} \\
&= -(\mathbf{I}_{m^2} + \mathbf{K}_{mm})[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \otimes \mathbf{M}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && (54)
\end{aligned}$$

Inverse Example 3: $\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X}^{-1}\mathbf{A}^T$

A third example: $\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X}^{-1}\mathbf{A}^T$ ($\mathbf{X}$ is symmetric).

If $\mathbf{X}$ is symmetric, then $\mathbf{X}^{-1}$ is symmetric and $\mathrm{d}(\mathbf{X})$ is symmetric.

Find the differential w.r.t. $\mathrm{d}\,\mathrm{vec}(\mathbf{X})$, but use the duplication matrix to limit the number of freely varying terms to those on the lower diagonal.

Inverse Example 3: $\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X}^{-1}\mathbf{A}^T$

To differentiate a symmetric matrix:

1. Take the full $\mathrm{d}\,\mathrm{vec}[\mathbf{F}(\mathbf{X})]$ differential.
2. Simplify as in every other differential.
3. After $\mathrm{d}\,\mathrm{vec}(\mathbf{X})$ is isolated, use the duplication matrix inside the differential operator to restrict $\mathrm{vec}(\mathbf{X}) = \mathbf{D}_n\,\mathrm{vech}(\mathbf{X})$.
4. Pull $\mathbf{D}_n$ outside of the differential operator because it is constant with respect to $\mathbf{X}$.

Inverse Example 3: $\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X}^{-1}\mathbf{A}^T$

First, taking differentials.

$$\begin{aligned}
\mathrm{d}[\mathbf{F}(\mathbf{X})] &= \mathrm{d}[\mathbf{A}\mathbf{X}^{-1}\mathbf{A}^T] \\
&= \mathbf{A}\,\mathrm{d}(\mathbf{X}^{-1})\mathbf{A}^T && \text{by (16)} \\
&= \mathbf{A}[-\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1}]\mathbf{A}^T && \text{by (50)} \\
&= -\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1}\mathbf{A}^T
\end{aligned}$$

Then vectorizing the differential.

$$\begin{aligned}
\mathrm{vec}[\mathrm{d}(\mathbf{F})] &= \mathrm{vec}[-\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1}\mathbf{A}^T] \\
&= -\,\mathrm{vec}[\mathbf{A}\mathbf{X}^{-1}\,\mathrm{d}(\mathbf{X})\mathbf{X}^{-1}\mathbf{A}^T] \\
&= -[(\mathbf{X}^{-1}\mathbf{A}^T)^T \otimes (\mathbf{A}\mathbf{X}^{-1})]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (4)}
\end{aligned}$$

Inverse Example 3: $\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X}^{-1}\mathbf{A}^T$

Continuing:

$$\begin{aligned}\text{vec}[\text{d}(\mathbf{F})] &= -[(\mathbf{X}^{-1}\mathbf{A}^T)^T \otimes (\mathbf{A}\mathbf{X}^{-1})]\,\text{d}\,\text{vec}(\mathbf{X}) \\ &= -[\mathbf{A}\mathbf{X}^{-1} \otimes \mathbf{A}\mathbf{X}^{-1}]\,\text{d}\,\text{vec}(\mathbf{X}) \qquad \text{by the Symmetry of } \mathbf{X}^{-1}\end{aligned}$$

We can finally impose the duplication identity.

$$\begin{aligned}\text{vec}[\text{d}(\mathbf{F})] &= -[\mathbf{A}\mathbf{X}^{-1} \otimes \mathbf{A}\mathbf{X}^{-1}]\,\text{d}[\text{vec}(\mathbf{X})] \qquad (55)\\ &= -[\mathbf{A}\mathbf{X}^{-1} \otimes \mathbf{A}\mathbf{X}^{-1}]\,\text{d}[\mathbf{D}_n\,\text{vech}(\mathbf{X})] \qquad (56)\\ &= -[\mathbf{A}\mathbf{X}^{-1} \otimes \mathbf{A}\mathbf{X}^{-1}]\mathbf{D}_n\,\text{d}\,\text{vech}(\mathbf{X}) \qquad (57)\end{aligned}$$

Inverse Example 4: $\varphi = \imath^T \mathbf{X}^{-1} \imath$

Assume $\imath$ is a vector of 1s. Then

$$\varphi = \imath^T \mathbf{X}^{-1} \imath$$

is the sum of **all** of the elements in $\mathbf{X}^{-1}$.

Now, if $\mathbf{X}$ is symmetric, then

$$\begin{aligned}
\mathrm{d}\,\mathrm{vec}(\varphi) &= \mathrm{vec}[\mathrm{d}(\imath^T \mathbf{X}^{-1} \imath)] && \\
&= \mathrm{vec}[\imath^T \,\mathrm{d}(\mathbf{X}^{-1}) \imath] && \text{by (16)} \\
&= -\,\mathrm{vec}[\imath^T \mathbf{X}^{-1} \,\mathrm{d}(\mathbf{X}) \mathbf{X}^{-1} \imath] && \text{by (50)} \\
&= -[(\mathbf{X}^{-1}\imath)^T \otimes \imath^T \mathbf{X}^{-1}]\,\mathrm{d}\,\mathrm{vec}(\mathbf{X}) && \text{by (4)} \\
&= -[\imath^T \mathbf{X}^{-1} \otimes \imath^T \mathbf{X}^{-1}]\mathbf{D}_n \,\mathrm{d}\,\mathrm{vech}(\mathbf{X}) && (58)
\end{aligned}$$

The Exponential: The Special Case

The Maclaurin Series is just the Taylor Series with the constant set to 0.

$$\varphi(x) = \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(0)}{k!} x^k$$

The definition of the Exponential Function:

(I) The derivative of e^x equals e^x.

(II) $\varphi^{(k)}(0) = 1$ for all k.

... which leads to the Maclaurin representation:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

We could also replace x with any function of x.

$$e^{f(x)} = \sum_{k=0}^{\infty} \frac{f(x)^k}{k!}$$

The Exponential: The Special Case

To find the differential of an exponential function, expand the function as a Maclaurin series and differentiate.

An example function for the exponential: $x\mathbf{A}$.

$$
\begin{aligned}
\mathrm{d}(e^{x\mathbf{A}}) = \mathrm{d}\left(\sum_{k=0}^{\infty}\frac{(x\mathbf{A})^k}{k!}\right) = \sum_{k=0}^{\infty}\mathrm{d}\left(\frac{x^k\mathbf{A}^k}{k!}\right) &= \sum_{k=0}^{\infty}\frac{\mathrm{d}(x^k\mathbf{A}^k)}{k!} \\
&= \sum_{k=0}^{\infty}\frac{\mathrm{d}(x^k)\mathbf{A}^k}{k!} \qquad \text{by (16)} \\
&= \sum_{k=0}^{\infty}\frac{[kx^{k-1}\,\mathrm{d}\,x]\mathbf{A}^k}{k!} \\
&= \sum_{k=0}^{\infty}\frac{x^{k-1}\mathbf{A}^k}{(k-1)!}\,\mathrm{d}\,x
\end{aligned}
$$

The Exponential: The Special Case

Factorials do not exist for negative numbers, so change the boundaries.

$$\begin{aligned} \mathrm{d}(e^{x\mathbf{A}}) = \sum_{k=0}^{\infty} \frac{x^{k-1}\mathbf{A}^{k}}{(k-1)!}\,\mathrm{d}\,x &= \sum_{k=1}^{\infty} \frac{x^{k-1}[\mathbf{A}\mathbf{A}^{k-1}]}{(k-1)!}\,\mathrm{d}\,x \\ &= \mathbf{A}\sum_{k=1}^{\infty} \frac{x^{k-1}\mathbf{A}^{k-1}}{(k-1)!}\,\mathrm{d}\,x \end{aligned}$$

And set $m = k - 1$:

$$\mathrm{d}(e^{x\mathbf{A}}) = \mathbf{A}\sum_{k=1}^{\infty} \frac{x^{k-1}\mathbf{A}^{k-1}}{(k-1)!}\,\mathrm{d}\,x = \mathbf{A}\sum_{m=0}^{\infty} \frac{x^{m}\mathbf{A}^{m}}{m!}\,\mathrm{d}\,x$$

Noticing that the summation is equal to the original exponenent:

$$\mathrm{d}(e^{x\mathbf{A}}) = \mathbf{A}\sum_{m=0}^{\infty} \frac{x^{m}\mathbf{A}^{m}}{m!}\,\mathrm{d}\,x = \mathbf{A}\sum_{m=0}^{\infty} \frac{(x\mathbf{A})^{m}}{m!}\,\mathrm{d}\,x = \mathbf{A}e^{x\mathbf{A}}\,\mathrm{d}\,x \tag{59}$$

The Exponential: The General Case

By analogy, define a matrix exponential.

$$\exp(\mathbf{X}) = \sum_{k=0}^{\infty}\left[\frac{1}{k!}\mathbf{X}^k\right] \tag{60}$$

To take the derivative of a matrix exponential, follow similar steps.

$$\begin{aligned}
\mathrm{d}\,\mathbf{F}(\mathbf{X}) &= \mathrm{d}[\exp(\mathbf{X})] \\
&= \mathrm{d}\left(\sum_{k=0}^{\infty}\left[\frac{1}{k!}\mathbf{X}^k\right]\right) \\
&= \sum_{k=0}^{\infty}\left[\frac{1}{k!}\,\mathrm{d}(\mathbf{X}^k)\right] && \text{by (16)} \\
&= \sum_{k=0}^{\infty}\left[\frac{1}{k!}\Big((\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-1} + \mathbf{X}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-2} + \cdots + \mathbf{X}^{k-1}(\mathrm{d}\,\mathbf{X})\Big)\right] \\
&&& \text{by (21)}
\end{aligned}$$

The Exponential: The General Case

Continuing:

$$
\begin{aligned}
\mathrm{d}\,\mathbf{F}(\mathbf{X}) &= \sum_{k=0}^{\infty}\left[\frac{1}{k!}\Big((\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-1} + \mathbf{X}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-2} + \cdots + \mathbf{X}^{k-1}(\mathrm{d}\,\mathbf{X})\Big)\right] \\
&= \sum_{k=0}^{\infty}\left[\frac{1}{k!}\left(\sum_{j=0}^{k-1}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-j-1}\right)\right)\right] \\
&= \sum_{k=1}^{\infty}\left[\frac{1}{k!}\sum_{j=0}^{k-1}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-j-1}\right)\right]
\end{aligned}
$$

Note that the bounds change because $0 \leq j \leq k-1$, but if $k = 0$, then $0 \leq j \leq 0 - 1 = -1$, which is a contradition.

The Exponential: The General Case

Setting $m = k - 1$, so $k = m + 1$.

$$\begin{aligned}
\mathrm{d}\,\mathbf{F}(\mathbf{X}) &= \sum_{k=1}^{\infty}\left[\frac{1}{k!}\sum_{j=0}^{k-1}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-j-1}\right)\right] \\
&= \sum_{m=0}^{\infty}\left[\frac{1}{(m+1)!}\sum_{j=0}^{m}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\right)\right]
\end{aligned}$$

Because everything until the $\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}$ is a scalar, we have

$$\begin{aligned}
\mathrm{tr}\left(\mathrm{d}[\exp(\mathbf{X})]\right) &= \mathrm{tr}\left(\sum_{m=0}^{\infty}\left[\frac{1}{(m+1)!}\sum_{j=0}^{m}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\right)\right]\right) \\
&= \sum_{m=0}^{\infty}\left[\frac{1}{(m+1)!}\sum_{j=0}^{m}\mathrm{tr}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\right)\right]
\end{aligned}$$

The Exponential: The General Case

Finishing:

$$
\begin{aligned}
\operatorname{tr}\Big(\mathrm{d}[\exp(\mathbf{X})]\Big) &= \sum_{m=0}^{\infty}\left[\frac{1}{(m+1)!}\sum_{j=0}^{m}\operatorname{tr}\Big(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\Big)\right] \\
&= \sum_{m=0}^{\infty}\left[\frac{1}{(m+1)!}(m+1)\ \operatorname{tr}\Big(\mathbf{X}^{m-j}\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\Big)\right] \\
&= \operatorname{tr}\left(\sum_{m=0}^{\infty}\left[\frac{1}{m!}\mathbf{X}^{m}\right](\mathrm{d}\,\mathbf{X})\right) \\
&= \operatorname{tr}\Big[\exp(\mathbf{X})\,\mathrm{d}(\mathbf{X})\Big] = \operatorname{vec}[\exp(\mathbf{X})^{T}]^{T}\,\mathrm{d}\operatorname{vec}(\mathbf{X}) \qquad (61)
\end{aligned}
$$

Note that we must take traces to obtain a sensible result.

The Logarithm: The Special Case

The Mercator Series is the Maclaurin Series Expansion for $\ln(1+x)$:

$$\begin{aligned}
\ln(1+x) &= \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(0)}{k!} x^k \\
&= \varphi(0) + \varphi'(0)x + \frac{\varphi''(0)x^2}{2!} + \frac{\varphi^{(3)}(0)x^3}{3!} + \frac{\varphi^{(4)}x^4(0)}{4!} + \dots \\
&= \ln(1+0) + \frac{x}{1+0} + (-1)\frac{x^2}{2!(1+0)^2} + (-2)(-1)\frac{x^3}{3!(1+0)^3} + \dots \\
&= 0 + x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}x^k}{k}
\end{aligned}$$

Replacing x with $-x$, all terms in the sum become negative.

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots = -\sum_{k=1}^{\infty} \frac{x^k}{k}$$

The Logarithm: The Special Case

To find the differential of a logarithmic function, expand as a Mercator series and differentiate.

An example function for the logarithm: $x\mathbf{A}$.

$$\begin{aligned}
\mathrm{d}\left[\ln(\mathbf{I}_n - x\mathbf{A})\right] = \mathrm{d}\left[-\sum_{k=1}^{\infty}\frac{(x\mathbf{A})^k}{k}\right] &= -\sum_{k=1}^{\infty}\frac{\mathrm{d}(x^k)\mathbf{A}^k}{k} && \text{by (16)} \\
&= -\sum_{k=1}^{\infty}\frac{kx^{k-1}\,\mathrm{d}\,x\mathbf{A}^k}{k} \\
&= -\mathbf{A}\sum_{k=1}^{\infty}[x^{k-1}\mathbf{A}^{k-1}]\,\mathrm{d}\,x \\
&= -\mathbf{A}\sum_{m=0}^{\infty}[x\mathbf{A}]^m\,\mathrm{d}\,x
\end{aligned}$$

The last line involved the changing-indices trick.

The Logarithm: The Special Case

And if $|x| < 1$ then, due to the geometric series,

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1 - x}$$

By analogy, if x and $\mathbf{A}$ satisfy a similar constraint, then

$$\begin{aligned} \mathrm{d}\left[\ln(\mathbf{I}_n - x\mathbf{A})\right] &= -\mathbf{A}\sum_{m=0}^{\infty}[x\mathbf{A}]^m \,\mathrm{d}\,x \\ &= -\mathbf{A}(\mathbf{I}_n - x\mathbf{A})^{-1}\,\mathrm{d}\,x \\ &= -\mathbf{A}(\mathbf{I}_n - x\mathbf{A})^{-1}\,\mathrm{d}\,x \qquad (62) \end{aligned}$$

The Logarithm: The General Case

For the multivariate case, define:

$$\ln(\mathbf{I}_n - \mathbf{X}) = -\sum_{k=1}^{\infty}\left[\frac{1}{k}\mathbf{X}^k\right]$$

To take the differential of $\ln(\mathbf{I}_n - \mathbf{X})$, notice that we ultimately use the same expansion as for the exponential differential.

$$\begin{aligned}
\mathrm{d}\,\mathbf{F}(\mathbf{X}) = \mathrm{d}\left[\ln(\mathbf{I}_n - \mathbf{X})\right] &= \mathrm{d}\left(-\sum_{k=1}^{\infty}\left[\frac{1}{k}\mathbf{X}^k\right]\right) \\
&= -\sum_{k=1}^{\infty}\left[\frac{1}{k}\,\mathrm{d}(\mathbf{X}^k)\right] && \text{by (16)} \\
&= -\sum_{k=1}^{\infty}\left[\frac{1}{k}\sum_{j=0}^{k-1}\left(\mathbf{X}^j(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-j-1}\right)\right]
\end{aligned}$$

The Logarithm: The General Case

Setting $m = k - 1$, so $k = m + 1$:

$$\begin{aligned} \mathrm{d}\,\mathbf{F}(\mathbf{X}) &= -\sum_{k=1}^{\infty}\left[\frac{1}{k}\sum_{j=0}^{k-1}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{k-j-1}\right)\right] \\ &= -\sum_{m=0}^{\infty}\left[\frac{1}{m+1}\sum_{j=0}^{m}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\right)\right] \end{aligned}$$

Because everything until the $\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}$ is a scalar, we have

$$\begin{aligned} \mathrm{tr}\left(\mathrm{d}\left[\ln(\mathbf{I}_n - \mathbf{X})\right]\right) &= \mathrm{tr}\left(-\sum_{m=0}^{\infty}\left[\frac{1}{m+1}\sum_{j=0}^{m}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\right)\right]\right) \\ &= -\sum_{m=0}^{\infty}\left[\frac{1}{m+1}\sum_{j=0}^{m}\mathrm{tr}\left(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\right)\right] \end{aligned}$$

The Logarithm: The General Case

Finishing:

$$\begin{aligned}
\operatorname{tr}\Big(\mathrm{d}\left[\ln(\mathbf{I}_n - \mathbf{X})\right]\Big) &= -\sum_{m=0}^{\infty}\left[\frac{1}{m+1}\sum_{j=0}^{m}\operatorname{tr}\Big(\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\mathbf{X}^{m-j}\Big)\right] \\
&= -\sum_{m=0}^{\infty}\left[\frac{1}{m+1}(m+1)\ \operatorname{tr}\Big(\mathbf{X}^{m-j}\mathbf{X}^{j}(\mathrm{d}\,\mathbf{X})\Big)\right] \\
&= -\operatorname{tr}\left(\sum_{m=0}^{\infty}\left[\mathbf{X}^{m}\right](\mathrm{d}\,\mathbf{X})\right) \\
&= -\operatorname{tr}\left[(\mathbf{I}_n - \mathbf{X})^{-1}\,\mathrm{d}\,\mathbf{X}\right] \\
&= -\operatorname{vec}\left[\left[(\mathbf{I}_n - \mathbf{X})^{-1}\right]^{T}\right]^{T}\mathrm{d}\operatorname{vec}(\mathbf{X}) \qquad (63)
\end{aligned}$$

Differentials of both exponentials and logarithms for multivariate functions behave in similar way to the univariate case but inside of trace operators.

References

- Abadir, K. M., & Magnus, J. R. (2005). *Matrix algebra*. Cambridge, UK: Cambridge University Press.
- Magnus, J. R. & Neudecker, H. (1986). Symmetry, 0-1 matrices and Jacobians: A review. *Econometric Theory*, *2*, 157–190.
- Magnus, J. R. & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and economics*. New York, NY: John Wiley & Sons