With(out) A Trace
Matrix Derivatives the Easy Way

Steven W. Nydick

University of Minnesota

May 16, 2012

Outline

Notation

- $\mathbf{A}$: A matrix
- $\mathbf{A}_c$: A matrix **held constant**

- $\mathbf{x}$: A vector
- y: A scalar (or a scalar 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 $\mathbf{X}^T$

- $\frac{\partial \mathbf{Y}}{\partial x}$: A matrix with elements $\frac{\partial y_{ij}}{\partial x}$
- $\frac{\partial y}{\partial \mathbf{X}}$: A matrix with elements $\frac{\partial y}{\partial x_{ij}}$

- $\langle \mathbf{x} \rangle_i$ or $\langle \mathbf{X} \rangle_{ij}$: The ith or ijth **place** of $\mathbf{x}$ or $\mathbf{X}$

Gradient, Jacobian, Hessian

A Gradient is the derivative of a scalar with respect to a vector.

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\left[\frac{\partial f(\mathbf{x})}{\partial x_1}\right] \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_2}\right] \quad \cdots \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_n}\right] \right)^T$$

If we have the function: $f(\mathbf{x}) = 2x_1x_2 + x_2^2 + x_1x_3^2$, then the Gradient is

$$\begin{aligned}\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} &= \left(\left[\frac{\partial f(\mathbf{x})}{\partial x_1}\right] \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_2}\right] \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_3}\right] \right)^T \\ &= \begin{bmatrix} 2x_2 + x_3^2 & 2x_1 + 2x_2 & 2x_1x_3 \end{bmatrix}^T\end{aligned}$$

Gradient, Jacobian, Hessian

A Jacobian is a the derivative of a vector with respect to a transposed vector.

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T} = \begin{pmatrix} \left[\frac{\partial f_1(\mathbf{x})}{\partial x_1}\right] & \cdots & \left[\frac{\partial f_1(\mathbf{x})}{\partial x_n}\right] \\ \vdots & \cdots & \vdots \\ \left[\frac{\partial f_k(\mathbf{x})}{\partial x_1}\right] & \cdots & \left[\frac{\partial f_k(\mathbf{x})}{\partial x_n}\right] \end{pmatrix}$$

If we have the function

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 3x_1^2 + x_2 & \ln(x_1) & \sin(x_2) \end{bmatrix}^T$$

Then the Jacobian is

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T} = \begin{pmatrix} 6x_1 & 1 \\ \frac{1}{x_1} & 0 \\ 0 & \cos(x_2) \end{pmatrix}$$

Gradient, Jacobian, Hessian

The Hessian is derivative of a Gradient with respect to a transposed vector.

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \begin{pmatrix} \left[\frac{\partial f(\mathbf{x})}{\partial x_1^2}\right] & \cdots & \left[\frac{\partial f(\mathbf{x})}{\partial x_1 \partial x_n}\right] \\ \vdots & \ddots & \vdots \\ \left[\frac{\partial f(\mathbf{x})}{\partial x_n \partial x_1}\right] & \cdots & \left[\frac{\partial f(\mathbf{x})}{\partial x_n^2}\right] \end{pmatrix}$$

Because our above Gradient is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} 2x_2 + x_3^2 & 2x_1 + 2x_2 & 2x_1 x_3 \end{bmatrix}^T$$

The Hessian would be

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \begin{pmatrix} 0 & 2 & 2x_3 \\ 2 & 2 & 0 \\ 2x_3 & 0 & 2x_1 \end{pmatrix}$$

Simplifying Classes of Matrix Derivatives

History of Schöneman's paper:

1. Wrote it while a post doc at UNC.
2. Originally submitted it to Psychometrika in 1965.
3. Editor mildly criticized paper.
    1. Compliment: reformulate certain problems (Lagrange multipliers) into interesting form (traces).
    2. Complaint: why would we want to do that?
4. Revised paper, resubmitted paper, but editorship changed hands, and took them almost a year to respond (asking for another revision).
    1. The new editor told him that a reviewer said: "nothing wrong with paper but not too important".

Simplifying Classes of Matrix Derivatives

History of Schöneman's paper:

5. Later learned that the original delay was caused by a statistician with expertise in matrix derivatives who thought that the paper would be published eventually.
6. The paper **was** published eventually ... **20 years later** in MBR.
7. Wrote the article "Better Never than Late: Peer Review and the Preservation of Prejudice" in 2001.

Simplifying Classes of Matrix Derivatives

There are two beneficial properties of Schöneman's paper:

1. Derivatives are always in matrix form.
2. No need for Dummy Matrices.

But, uses traces, and thus, uses trace properties.

So ... A Review of Traces/Trace Properties:

What is a Trace?

Definition:

$$\text{tr}\left(\mathbf{Y}\right) = \sum_i (y_{ii}), \quad \mathbf{Y} \text{ is square}$$

OK - that's simple, but what does that mean?

Well, take a square matrix and add up the diagonal elements

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, \quad \text{tr}\left(\mathbf{A}\right) = a_{11} + a_{22} + \cdots + a_{nn}$$

Linearity of Traces

The **MOST** important aspect of traces (for our later derivations):

$$\text{tr} : M(\mathbb{R})^n \to \mathbb{R}^1 \quad \text{is linear}$$

Thus:

$$\text{tr}\left(\mathbf{A} + \mathbf{B}\right) = \text{tr}\left(\mathbf{A}\right) + \text{tr}\left(\mathbf{B}\right) \tag{1}$$

and

$$\text{tr}\left(c\mathbf{A}\right) = c \; \text{tr}\left(\mathbf{A}\right) \tag{2}$$

Transposition of Dependent Variable

Traces have **SEVERAL OTHER** important properties.

Property 1: Transposition of Dependent Variable

We have:

$$\mathrm{tr}\left(\mathbf{Y}\right) = \mathrm{tr}\left(\mathbf{Y}^T\right) \tag{3}$$

Thus:

$$\frac{\partial \mathrm{tr}\left(\mathbf{Y}\right)}{\partial \mathbf{X}} = \frac{\partial \mathrm{tr}\left(\mathbf{Y}^T\right)}{\partial \mathbf{X}}$$

Cyclic Permutation

Property 2: Cyclic Permutation

We have:

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}) \tag{4}$$

Why? Well, start from the left of Equation (4).

$$\begin{aligned}
\text{tr}(\mathbf{AB}) &= \text{tr}\left[\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}\begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix}\right] \\
&= a_{11}b_{11} + \cdots + a_{1m}b_{m1} + \sum_{i=1}^{m}(a_{2i}b_{i2}) + \cdots + \sum_{i=1}^{m}(a_{ni}b_{in}) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{m}(a_{ji}b_{ij})
\end{aligned}$$

Cyclic Permutation

And also start from the right of Equation (4).

$$\begin{aligned}
\text{tr}(\mathbf{BA}) &= \text{tr}\left[\begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix}\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}\right] \\
&= \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n}(b_{ij}a_{ji}) = \sum_{j=1}^{n}\sum_{i=1}^{m}(b_{ij}a_{ji}) \quad = \displaystyle\sum_{j=1}^{n}\sum_{i=1}^{m}(a_{ji}b_{ij}) \\
&= \text{tr}(\mathbf{AB})
\end{aligned}$$

So:

$$\frac{\partial \text{tr}(\mathbf{AB})}{\partial \mathbf{X}} = \frac{\partial \text{tr}(\mathbf{BA})}{\partial \mathbf{X}}$$

Rotating the order **does not** change the trace of square matrices.

Cyclic Permutation

A consequence of the last derivation:

- Let $\mathbf{U}$ and $\mathbf{H}$ have the same dimensions.
- If you want to multiply paired entries (e.g. $u_{ij}h_{ij}$) and add all the multiplications:
    - Flip one of the matrices, multiply, and take the trace of that multiplication.

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(u_{ij}h_{ij}) = \sum_{j=1}^{m}\sum_{i=1}^{n}\left((u^T)_{ji}h_{ij}\right) = \operatorname{tr}\left(\mathbf{U}^T\mathbf{H}\right) \tag{5}$$

Transposition of Independent Variable

Calculus Property 1: Transposition of Independent Variable

By definition:

$$\frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial \mathbf{X}} = \frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial x_{ij}}, \, i = 1, \ldots, n; \, j = 1, \ldots, m$$

where $\frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial x_{ij}}$ is what we put in the ijth place in our derivative matrix.

Thus:

$$\frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial (\mathbf{X}^T)} = \frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial x_{ji}}, \, (j = 1, \ldots, m; \, i = 1, \ldots, n) = \left(\frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial \mathbf{X}}\right)^T \quad (6)$$

because $\frac{\partial \text{tr}\left(\mathbf{Y}\right)}{\partial x_{ji}}$ is what we put in the ijth place in our derivative matrix.

Transposition of Independent Variable

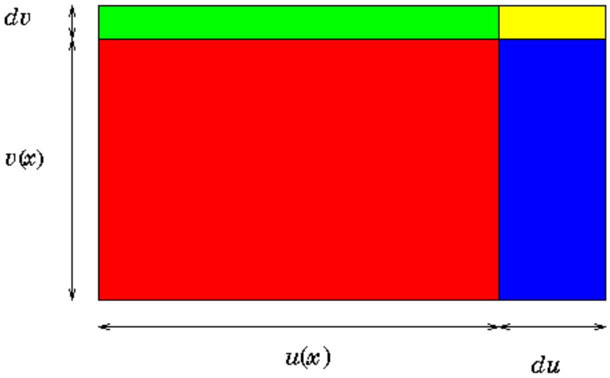
Deriving with respect to a transposed variable replaces each entry in the new matrix with the **derivative** of the **corresponding transposed component**.

Replacing every entry with the derivative of the transposed component $\rightarrow$ Transposing the entire matrix of partial derivatives.

Product Rule

Calculus Property 2: Product Rule

An illustration of the product rule:



Product Rule

Calculus Property 2: Product Rule

Based on the previous illustration:

$$\frac{d(uv)}{dx} = \left(\frac{du}{dx}\right)(v) + (u)\left(\frac{dv}{dx}\right)$$

In this case, u and v are scalar **functions** of x.

Now, we want to translate this to matrices and traces of matrices:

Pick any row $\mathbf{u}_i$ and any column $\mathbf{v}_j$ from $\mathbf{U}$ and $\mathbf{V}$.

Product Rule

If we take the derivative of the matrix product with respect to a scalar:

$$\frac{\partial(\mathbf{U}\mathbf{V})}{\partial x}$$

we find that the i,jth place in our new derivative matrix is

$$\begin{aligned}
\frac{\partial(\mathbf{u}_{i.}^T\mathbf{v}_{.j})}{\partial x} &= \frac{\partial(u_{i1}v_{1j} + \cdots + u_{in}v_{nj})}{\partial x} \\
&= \frac{\partial(u_{i1}v_{1j})}{\partial x} + \cdots + \frac{\partial(u_{in}v_{nj})}{\partial x} \\
&= \frac{\partial u_{i1}}{\partial x}v_{1j} + u_{i1}\frac{\partial v_{1j}}{\partial x} + \cdots + \frac{\partial u_{in}}{\partial x}v_{nj} + u_{in}\frac{\partial v_{nj}}{\partial x}
\end{aligned}$$

Product Rule

So, now we want to collect terms:

$$\begin{aligned}\frac{\partial(\mathbf{u}_{i.}^T\mathbf{v}_{.j})}{\partial x} &= \frac{\partial u_{i1}}{\partial x}v_{1j} + u_{i1}\frac{\partial v_{1j}}{\partial x} + \cdots + \frac{\partial u_{in}}{\partial x}v_{nj} + u_{in}\frac{\partial v_{nj}}{\partial x} \\ &= \left(\frac{\partial u_{i1}}{\partial x}v_{1j} + \cdots + \frac{\partial u_{in}}{\partial x}v_{nj}\right) + \left(u_{i1}\frac{\partial v_{1j}}{\partial x} + \cdots + u_{in}\frac{\partial v_{nj}}{\partial x}\right) \\ &= \frac{\partial \mathbf{u}_{i.}^T}{\partial x}\mathbf{v}_{.j} + \mathbf{u}_{i.}^T\frac{\partial \mathbf{v}_{.j}}{\partial x}\end{aligned}$$

And because our element is **arbitrary**, we can generalize:

$$\begin{aligned}\frac{\partial(\mathbf{UV})}{\partial x} &= \frac{\partial \mathbf{U}}{\partial x}\mathbf{V} + \mathbf{U}\frac{\partial \mathbf{V}}{\partial x} \\ &= \frac{\partial(\mathbf{UV}_c)}{\partial x} + \frac{\partial(\mathbf{U}_c\mathbf{V})}{\partial x} \qquad (7)\end{aligned}$$

Product Rule

There are two notes on the product rule:

Note 1: For the product rule to make sense, both **U** and **V** should be functions of **X**.

For the Univariate Case, let $u = x^2 + 2$ and $v = 2x + \sin(x)$. Then:

$$
\begin{aligned}
\frac{d(uv)}{dx} &= \left(\frac{du}{dx}\right)(v) + (u)\left(\frac{dv}{dx}\right) \\
&= (2x)\Big(2x + \sin(x)\Big) + (x^2 + 2)\Big(2 + \cos(x)\Big) \\
&= 4x^2 + 2x\sin(x) + 2x^2 + 4 + x^2\cos(x) + 2\cos(x) \\
&= x^2\Big(6 + \cos(x)\Big) + 2\Big(x\sin(x) + \cos(x)\Big) + 4
\end{aligned}
$$

We will discuss the multivariate case later.

Product Rule

There are two notes on the product rule:

Note 2: We can put $\mathbf{V}_c$ and $\mathbf{U}_c$ inside the derivative funtion, but they are now **constants** with respect to $\mathbf{X}$, even if they are functions of $\mathbf{X}$.

For the Univariate Case, let $u = x^3 + \ln(x)$ and $v = 3x^2$. Then:

$$\begin{aligned}\frac{d(uv_c)}{dx} &= \frac{d\left[\left(x^3 + \ln(x)\right)(3x^2)_c\right]}{dx} \\ &= \left(3x^2 + \frac{1}{x}\right)(3x^2) \\ &= 9x^4 + 3x\end{aligned}$$

We will discuss the multivariate case later.

Multidimensional Chain Rule

Calculus Property 3: Chain Rule

Let:

$$z = 2x_1^2 + x_1 \cos(x_2)$$

Then, by definition:

$$\frac{\partial z}{\partial \mathbf{x}} = \begin{pmatrix} \left[\frac{\partial z}{\partial x_1}\right] \\ \left[\frac{\partial z}{\partial x_2}\right] \end{pmatrix} = \begin{pmatrix} 4x_1 + \cos(x_2) \\ \\ -x_1 \sin(x_2) \end{pmatrix}$$

Our partial derivative with respect to x_1 is $4x_1 + \cos(x_2)$, and our partial derivative with respect to x_2 is $-x \sin(x_2)$. Furthermore, these go in the respective parts of our derivative matrix (replacing x_1 and x_2).

Multidimensional Chain Rule

Now if:

$$x_1 = 3t \qquad \text{and} \qquad x_2 = t$$

Then:

$\frac{dz}{dt}$ is now the derivative with respect to t accounted for by x_1 **and** the derivative with respect to t accounted for by x_2.

And we account for:

$$[4(3t) + \cos(t)]\frac{\partial x_1}{\partial t} \qquad \text{by } x_1$$
$$-(3t)\sin(t)\frac{\partial x_2}{\partial t} \qquad \text{by } x_2$$

Multidimensional Chain Rule

Thus:

$$\frac{dz}{dt} = \left(\frac{\partial z}{\partial \mathbf{x}}\right)^T \frac{\partial \mathbf{x}}{\partial t} = \sum_{i=1}^{2} \left(\frac{\partial z}{\partial x_i} \frac{\partial x_i}{\partial t}\right)$$
$$= [12t + \cos(t)](3) + [-(3t)\sin(t)](1)$$
$$= 36t + 3\cos(t) - 3t\sin(t)$$

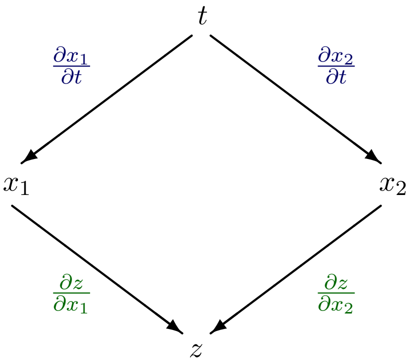
But: $z = 2x_1^2 + x_1 \cos(x_2) = 2(3t)^2 + (3t)\cos(t) = 18t^2 + 3t\cos(t)$

So, another way of getting the same result:

$$\frac{dz}{dt} = \frac{d[18t^2 + 3t\cos(t)]}{dt}$$
$$= 36t + 3t[-\sin(t)] + 3\cos(t) = 36t + 3\cos(t) - 3t\sin(t)$$

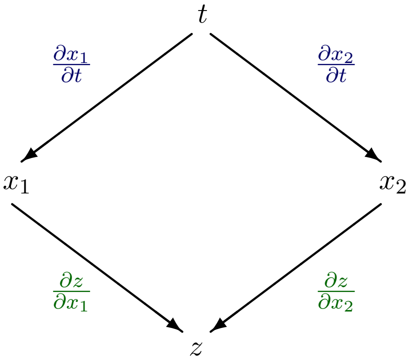
An Easy Way to Remember the Chain Rule

Because effects are slopes and slopes are derivatives, writing out a path diagram from t to z would have the derivatives along the paths.



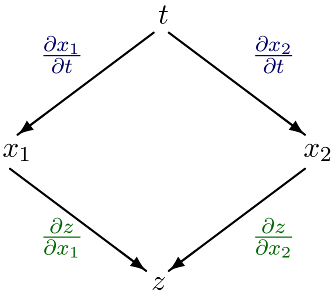
The **total effect** of t on z is found by multiplying the effects down each path and summing the total effects across paths.

An Easy Way to Remember the Chain Rule



For example, let's find the total effect of a 1 unit change in t on z. Well, if t changes 1 unit, then x_1 changes $\frac{dx_1}{dt}$ units (because the derivative is the slope of t on x_1). Moreover, if x_1 changes 1 unit, then z changes $\frac{dz}{dx_1}$ units (because the derivative is the slope of x_1 on z).

An Easy Way to Remember the Chain Rule



Therefore, if t changes 1 unit, then it's effect on z through x_1 would be the distance it travels in the x_1 direction:

$$x_1 \text{ distance} = \frac{dt}{dx_1} \times 1 = \frac{dt}{dx_1}$$

multiplied by how much a unit change in the x_1 direction changes z:

$$z \text{ distance through } x_1 = \frac{dx_1}{dz} \times (x_1 \text{ distance}) = \frac{dx_1}{dz}\frac{dt}{dx_1}$$

An Easy Way to Remember the Chain Rule

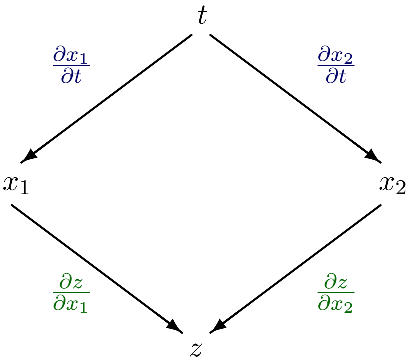
And if t changes 1 unit, then it's effect on z through x_2 would be:

$$z \text{ distance through } x_2 = \frac{dx_2}{dz} \times (x_2 \text{ distance}) = \frac{dx_2}{dz}\frac{dt}{dx_2}$$

Thus, if t moves 1 unit, it moves z: $\left(\frac{dx_1}{dz}\frac{dt}{dx_1}\right)$ through x_1 **and** it moves z: $\left(\frac{dx_2}{dz}\frac{dt}{dx_2}\right)$ through x_2, so it **in total** moves z:

$$z \text{ total distance} = \frac{dx_1}{dz}\frac{dt}{dx_1} + \frac{dx_2}{dz}\frac{dt}{dx_2}$$

An Easy Way to Remember the Chain Rule



Or, as written before, to find the **effect** of t on z:

$$\frac{dz}{dt} = \left(\frac{\partial z}{\partial x_1}\right)\left(\frac{\partial x_1}{\partial t}\right) + \left(\frac{\partial z}{\partial x_2}\right)\left(\frac{\partial x_2}{\partial t}\right) = \sum_{i=1}^{2}\left(\frac{\partial z}{\partial x_i}\frac{\partial x_i}{\partial t}\right)$$

Matrices Chain Rule

And because as our vector chain rule:

$$\frac{d[f(\mathbf{y})]}{dx} = \sum_i \left(\frac{d[f(\mathbf{y})]}{d[y_i(x)]} \frac{d[y_i(x)]}{dx} \right) \tag{8}$$

We can expand upon that to obtain a chain rule for matrices.

$$\frac{\partial[f(\mathbf{Y})]}{\partial x_{pq}} = \sum_i \sum_j \left(\frac{\partial[f(\mathbf{Y})]}{\partial[y_{ij}(x_{pq})]} \frac{\partial[y_{ij}(x_{pq})]}{\partial x_{pq}} \right) \tag{9}$$

In Equation (9) y_{ij} is a function of x_{pq}, and we have to take the derivative with respect to each of the elements in $\mathbf{Y}$.

Derivatives

Here is the standard derivative definition:

$$Df(x) = \lim_{t \to 0} \frac{f(x+t) - f(t)}{t}$$

The equation is a infinitesimal form of $m = \frac{\Delta y}{\Delta x}$; it is finding the slope or **linear approximation** to this function as the distance between the points on the x-axis goes to 0.

If there is a large distance between points on the x-axis, and if the function is not linear, then the slope will not be a good representation of how the function is changing. However, as the distance between points on the x-axis goes to 0, the mini function becomes more linear.

Directional Derivatives: Vectors

In vector calculus, there is a similar equation.

$$D_{\mathbf{w}}f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{w}) - f(\mathbf{x})}{t}$$

Now, our function is a surface (a scalar function of as many dimensional inputs as there are elements in $\mathbf{x}$), so the derivative will change at a multidimensional point $\mathbf{x}$ based on the direction we travel from that point.

Think of what happens if you were to stand on a mountain and turn around in a circle: in some directions, the slope will be very steep (and you might fall off the mountain), but in other directions, there will barely be any slope at all.

Directional Derivatives: Vectors

Now **w** tells us which direction we want to be facing when we calculate the derivative at a specific point **x**.

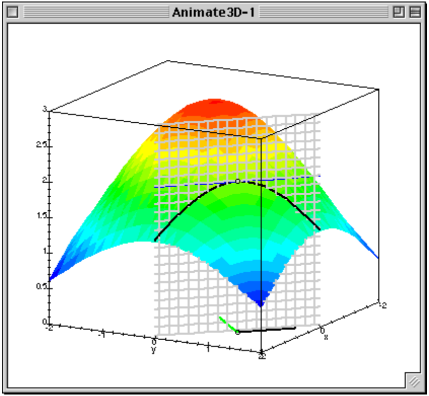
$$D_{\mathbf{w}}f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{w}) - f(\mathbf{x})}{t}$$

The directional derivative is basically telling us what is the best **linear approximation** of this function at a particular point if we are facing up the mountain, down the mountain, at a 45 degree angle up the mountain, etc.

Directional Derivatives: Vectors

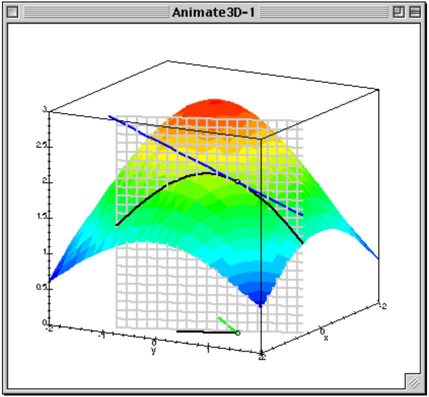
If our **x** vector is **two** dimensional, then the function would form a mountain in **three** dimensional space.

One Direction (at a given point):



Directional Derivatives: Vectors

A Second Direction (at the same point):



Notice how the steepness of the slope changes at both points.

Directional Derivatives: Vectors

First -- pick an abritrary unit length $\mathbf{w}$:

$$\mathbf{w}^T\mathbf{w} = 1$$

Second -- set up the standard, directional derivative definition:

$$D_{\mathbf{w}}f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{w}) - f(\mathbf{x})}{t}$$

If $\mathbf{w}_{(i)} = (0, 0, 0, 0, 1, 0, \ldots, 0)$ where 1 is in $\langle \mathbf{w} \rangle_i$, then

$$D_{\mathbf{w}_{(i)}}f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{w}_{(i)}) - f(\mathbf{x})}{t}$$

will reduce to the **regular** partial derivative in the ith place.

Directional Derivatives: Vectors

Now, if we can find a $\mathbf{u}$, such that:

$$D_{\mathbf{w}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{w}) - f(\mathbf{x})}{t} = \mathbf{w}^T \mathbf{u}$$

Then, for an arbitrary place i, in an arbitrary direction $\langle \mathbf{w} \rangle_i$:

$$D_{\mathbf{w}_{(i)}} f(\mathbf{x}) = \mathbf{w}_{(i)}^T \mathbf{u} \qquad \underline{\text{reduces to}} \qquad D_{\mathbf{w}_{(i)}} f(\mathbf{x}) = u_i$$

where u_i is the partial derivative in the ith place.

Because the ith place is arbitrary:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{u}$$

Directional Derivatives: Matrices

Now let $\mathbf{Y}_{(ij)}$ be a Matrix such that $y_{ij} = 1$ in $\langle \mathbf{Y} \rangle_{ij}$ and 0 elsewhere.

Then, extending our directional derivative definition to matrices:

$$D_{\mathbf{Y}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t} \tag{10}$$

We can conclude that

$$D_{\mathbf{Y}_{(ij)}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}_{(ij)}) - f(\mathbf{X})}{t}$$

will "pick off" the partial derivative in the ijth place.

Directional Derivatives: Matrices

Now, if we can find a $\mathbf{U}$, such that

$$D_{\mathbf{Y}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t} = \operatorname{tr}\left(\mathbf{Y}^T \mathbf{U}\right) \tag{11}$$

Then, for an arbitrary place ij, in an arbitrary direction $\langle \mathbf{Y} \rangle_{ij}$

$$\begin{aligned} D_{\mathbf{Y}_{(ij)}} f(\mathbf{X}) = \operatorname{tr}(\mathbf{Y}_{(ij)}^T \mathbf{U}) &= \sum_j \sum_i (y_{ij} u_{ij}) \qquad \text{by (5)} \\ &= u_{ij} \end{aligned}$$

Because the ijth place is arbitrary:

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{U} \tag{12}$$

Directional Derivatives: Matrices

First -- put in the form of the definition:

$$D_{\mathbf{Y}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t}$$

Second -- simplify until you can find the equality:

$$D_{\mathbf{Y}} f(\mathbf{X}) = \mathrm{tr}\left(\mathbf{Y}^T \mathbf{U}\right)$$

Third -- remove your $\mathbf{U}$, and note that:

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{U}$$

Definition

Our 1st function:

$$f(\mathbf{X}) = \text{tr}(\mathbf{AX})$$

Our objective is to find:

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial\, \text{tr}(\mathbf{AX})}{\partial \mathbf{X}}$$

Only by simplifying the definition:

$$D_{\mathbf{Y}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t}$$

Calculation

$$
\begin{aligned}
D_{\mathbf{Y}} f(\mathbf{X}) &= \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t} && \text{by (10)} \\
&= \lim_{t \to 0} \frac{\mathrm{tr}(\mathbf{A}[\mathbf{X} + t\mathbf{Y}]) - \mathrm{tr}(\mathbf{AX})}{t} \\
&= \lim_{t \to 0} \frac{\mathrm{tr}(\mathbf{AX} + \mathbf{A}t\mathbf{Y}) - \mathrm{tr}(\mathbf{AX})}{t} \\
&= \lim_{t \to 0} \frac{\mathrm{tr}(t\mathbf{AY})}{t} && \text{by (1)} \\
&= \lim_{t \to 0} \mathrm{tr}(\mathbf{AY}) && \text{by (2)} \\
&= \mathrm{tr}(\mathbf{AY}) \\
&= \mathrm{tr}([\mathbf{AY}]^T) && \text{by (3)} \\
&= \mathrm{tr}(\mathbf{Y}^T \mathbf{A}^T)
\end{aligned}
$$

Result

So, we found that:

$$D_{\mathbf{Y}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t}$$
$$= \operatorname{tr}(\mathbf{Y}^T \mathbf{A}^T) = \operatorname{tr}(\mathbf{Y}^T \mathbf{U}) \qquad \text{by (11)}$$

And we can spot that in **this** case:

$$\mathbf{U} = \mathbf{A}^T$$

And thus, by Equation (12):

$$\mathbf{U} = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$
$$= \frac{\partial \operatorname{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}^T \tag{13}$$

Definition

Our 2nd function:

$$f(\mathbf{X}) = \text{tr}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})$$

Our objective is to find:

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial\, \text{tr}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}}$$

Only by simplifying the definition:

$$D_{\mathbf{Y}} f(\mathbf{X}) = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t}$$

Calculation

$$
\begin{aligned}
D_{\mathbf{Y}} f(\mathbf{X}) &= \lim_{t\to 0} \frac{f(\mathbf{X}+t\mathbf{Y}) - f(\mathbf{X})}{t} && \text{by (10)} \\
&= \lim_{t\to 0} \frac{\mathrm{tr}([\mathbf{X}+t\mathbf{Y}]^T\mathbf{A}[\mathbf{X}+t\mathbf{Y}]\mathbf{B}) - \mathrm{tr}(\mathbf{X}^T\mathbf{AXB})}{t} \\
&= \lim_{t\to 0} \frac{\mathrm{tr}([\mathbf{X}+t\mathbf{Y}]^T\mathbf{A}[\mathbf{X}+t\mathbf{Y}]\mathbf{B} - \mathbf{X}^T\mathbf{AXB})}{t} && \text{by (1)} \\
&= \lim_{t\to 0} \frac{\mathrm{tr}(\mathbf{X}^T\mathbf{A}t\mathbf{YB} + t\mathbf{Y}^T\mathbf{AXB} + t\mathbf{Y}^T\mathbf{A}t\mathbf{YB})}{t} \\
&= \lim_{t\to 0} \frac{\mathrm{tr}(t[\mathbf{X}^T\mathbf{AYB} + \mathbf{Y}^T\mathbf{AXB} + t\mathbf{Y}^T\mathbf{AYB}])}{t} \\
&= \lim_{t\to 0}[\mathrm{tr}(\mathbf{X}^T\mathbf{AYB} + \mathbf{Y}^T\mathbf{AXB} + t\mathbf{Y}^T\mathbf{AYB})] && \text{by (2)}
\end{aligned}
$$

Calculation

Continuing:

$$
\begin{aligned}
D_{\mathbf{Y}}f(\mathbf{X}) &= \lim_{t\to 0}[\mathrm{tr}(\mathbf{X}^T\mathbf{AYB} + \mathbf{Y}^T\mathbf{AXB} + t\mathbf{Y}^T\mathbf{AYB})] \\
&= \lim_{t\to 0}[\mathrm{tr}(\mathbf{X}^T\mathbf{AYB} + \mathbf{Y}^T\mathbf{AXB})] + \lim_{t\to 0}[t\ \mathrm{tr}(\mathbf{Y}^T\mathbf{AYB})] \\
&\qquad \text{by (1) \& (2)} \\
&= \lim_{t\to 0}[\mathrm{tr}(\mathbf{X}^T\mathbf{AYB} + \mathbf{Y}^T\mathbf{AXB})] \\
&= \mathrm{tr}(\mathbf{X}^T\mathbf{AYB} + \mathbf{Y}^T\mathbf{AXB}) \\
&= \mathrm{tr}(\mathbf{X}^T\mathbf{AYB}) + \mathrm{tr}(\mathbf{Y}^T\mathbf{AXB}) && \text{by (1)} \\
&= \mathrm{tr}(\mathbf{BX}^T\mathbf{AY}) + \mathrm{tr}(\mathbf{Y}^T\mathbf{AXB}) && \text{by (4)}
\end{aligned}
$$

Calculation

And Finally:

$$
\begin{aligned}
D_{\mathbf{Y}}f(\mathbf{X}) &= \mathrm{tr}(\mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{Y}) + \mathrm{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{X}\mathbf{B}) \\
&= \mathrm{tr}[(\mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{Y})^T] + \mathrm{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{X}\mathbf{B}) && \text{by (3)} \\
&= \mathrm{tr}(\mathbf{Y}^T\mathbf{A}^T\mathbf{X}\mathbf{B}^T) + \mathrm{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{X}\mathbf{B}) \\
&= \mathrm{tr}(\mathbf{Y}^T\mathbf{A}^T\mathbf{X}\mathbf{B}^T + \mathbf{Y}^T\mathbf{A}\mathbf{X}\mathbf{B}) && \text{by (1)} \\
&= \mathrm{tr}(\mathbf{Y}^T[\mathbf{A}^T\mathbf{X}\mathbf{B}^T + \mathbf{A}\mathbf{X}\mathbf{B}])
\end{aligned}
$$

Result

So, we found that:

$$\begin{aligned} D_{\mathbf{Y}} f(\mathbf{X}) &= \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{Y}) - f(\mathbf{X})}{t} \\ &= \mathrm{tr}(\mathbf{Y}^T [\mathbf{A}^T \mathbf{X} \mathbf{B}^T + \mathbf{A}\mathbf{X}\mathbf{B}]) = \mathrm{tr}(\mathbf{Y}^T \mathbf{U}) \qquad \text{by (11)} \end{aligned}$$

And we can spot that in **this** case:

$$\mathbf{U} = \mathbf{A}^T \mathbf{X} \mathbf{B}^T + \mathbf{A}\mathbf{X}\mathbf{B}$$

And thus, by Equation (12):

$$\begin{aligned} \mathbf{U} &= \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \\ &= \frac{\partial \, \mathrm{tr}(\mathbf{X}^T \mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}} = \mathbf{A}^T \mathbf{X} \mathbf{B}^T + \mathbf{A}\mathbf{X}\mathbf{B} \end{aligned} \tag{14}$$

Definition

Our 3rd function, assuming that $\mathbf{Y}$ is non-singular and depends on $\mathbf{X}$:

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

Our objective is to find a better expression for

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial\, \text{tr}(\mathbf{Y}^{-1})}{\partial \mathbf{X}}$$

by working with previous trace derivative rules.

Calculation

We have:

$$\frac{\partial\,\text{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}(\mathbf{Y}^{-2}\mathbf{Y})}{\partial\mathbf{X}}$$

$$= \frac{\partial\,\text{tr}(\mathbf{Y}_c^{-2}\mathbf{Y})}{\partial\mathbf{X}} + \boxed{\frac{\partial\,\text{tr}(\mathbf{Y}^{-2}\mathbf{Y}_c)}{\partial\mathbf{X}}} \qquad \text{by (7)}$$

Zoom In

$$\boxed{\frac{\partial\,\text{tr}(\mathbf{Y}^{-1}\mathbf{Y}^{-1}\mathbf{Y}_c)}{\partial\mathbf{X}}} = \frac{\partial\,\text{tr}(\mathbf{Y}_c\mathbf{Y}_c^{-1}\mathbf{Y}^{-1})}{\partial\mathbf{X}} + \frac{\partial\,\text{tr}(\mathbf{Y}^{-1}\mathbf{Y}_c^{-1}\mathbf{Y}_c)}{\partial\mathbf{X}} \qquad \text{by (7) \& (4)}$$

$$= \frac{\partial\,\text{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} + \frac{\partial\,\text{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} = \frac{2\partial\,\text{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} \qquad (15)$$

Result

Therefore:

$$\begin{aligned}
\frac{\partial\,\mathrm{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} &= \frac{\partial\,\mathrm{tr}(\mathbf{Y}_c^{-2}\mathbf{Y})}{\partial\mathbf{X}} + \frac{\partial\,\mathrm{tr}(\mathbf{Y}^{-2}\mathbf{Y}_c)}{\partial\mathbf{X}} \\
\frac{\partial\,\mathrm{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} &= \frac{\partial\,\mathrm{tr}(\mathbf{Y}_c^{-2}\mathbf{Y})}{\partial\mathbf{X}} + \frac{2\partial\,\mathrm{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} && \text{by (15)} \\
-\frac{\partial\,\mathrm{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} &= \frac{\partial\,\mathrm{tr}(\mathbf{Y}_c^{-2}\mathbf{Y})}{\partial\mathbf{X}}
\end{aligned}$$

And, finally, after multiplying by (-1) on both sides:

$$\frac{\partial\,\mathrm{tr}(\mathbf{Y}^{-1})}{\partial\mathbf{X}} = -\frac{\partial\,\mathrm{tr}(\mathbf{Y}_c^{-2}\mathbf{Y})}{\partial\mathbf{X}} \tag{16}$$

We have turned a “derivative of the trace-inverse” problem into a standard “trace derivative” problem.

Definition

Our 4th function, assuming that $\mathbf{Y}$ is non-singular and depends on $\mathbf{X}$:

$$f(\mathbf{X}) = |\mathbf{Y}|$$

Our objective is to find a better expression for

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial |\mathbf{Y}|}{\partial \mathbf{X}}$$

by working with previous trace derivative rules and determinant rules.

Determinant Review

We know from linear algebra:

$$\mathbf{Y}^{-1} = \frac{1}{|\mathbf{Y}|}\mathbf{Q} \qquad \text{where } \mathbf{Q} \text{ is the adjoint matrix}$$

Process for calculating $(q^T)_{ij}$:

1. Cross out row i and column j of $\mathbf{Y}$.
2. Take determinant of the smaller $[(n-1) \times (n-1)]$ matrix.
3. If $i + j$ is odd, then negate the previous step.

Thus:

$$|\mathbf{Y}|\mathbf{I} = \mathbf{QY} \tag{17}$$

Adjoint Review

- Note: $(q^T)_{ij} = q_{ji}$ does not depend on any of the elements in row i **or** column j of $\mathbf{Y}$.
- Thus, q_{ji} does not depend on y_{ij}.
- So: $\frac{\partial(q_{ji}y_{ij})}{\partial y_{ij}} = q_{ji}$

Based on the previous slide we have:

$$|\mathbf{Y}|\mathbf{I} = \mathbf{QY} \qquad \text{by (17)}$$

$$\begin{pmatrix} |\mathbf{Y}| & 0 & \cdots & 0 \\ 0 & |\mathbf{Y}| & \vdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & |\mathbf{Y}| \end{pmatrix} = \begin{pmatrix} \sum_p (q_{1p}y_{p1}) & & & \mathbf{O} \\ & \ddots & & \\ & & \ddots & \\ \mathbf{O} & & & \sum_p (q_{np}y_{pn}) \end{pmatrix}$$

Determinant Derivative, Part 1

Thus, given **any** j such that $1 \leq j \leq n$:

$$|\mathbf{Y}| = \sum_p (q_{jp} y_{pj}) \tag{18}$$

And, for an arbitrary y_{ij}, pick the jth row of $\mathbf{q}$ and column of $\mathbf{y}$:

$$\frac{\partial|\mathbf{Y}|}{\partial y_{ij}} = \frac{\partial\left(\sum_p (q_{jp} y_{pj})\right)}{\partial y_{ij}} \qquad \text{by (18)}$$

$$= \sum_p \left(q_{jp} \frac{\partial y_{pj}}{\partial y_{ij}}\right) = q_{ji} \frac{\partial y_{ij}}{\partial y_{ij}} = q_{ji} = (q^T)_{ij} \tag{19}$$

Because y_{ij} was arbitrary, for an entire matrix:

$$\frac{\partial|\mathbf{Y}|}{\partial \mathbf{Y}} = \mathbf{Q}^T \tag{20}$$

Determinant Derivative, Part 2

Now, for an arbitrary pqth element of $\mathbf{X}$ (where $\mathbf{Y}$ depends on $\mathbf{X}$):

$$\begin{aligned}
\frac{\partial|\mathbf{Y}|}{\partial x_{pq}} &= \sum_i \sum_j \left(\frac{\partial|\mathbf{Y}|}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{pq}}\right) && \text{by (9)} \\
&= \sum_i \sum_j \left(q_{ji} \frac{\partial y_{ij}}{\partial x_{pq}}\right) && \text{by (19)} \\
&= \frac{\partial\left(\sum_i \sum_j (q_{\mathbf{c}ji} y_{ij})\right)}{\partial x_{pq}} \\
&= \frac{\partial \operatorname{tr}(\mathbf{Q}_c \mathbf{Y})}{\partial x_{pq}} && \text{by (5)}
\end{aligned}$$

Because x_{pq} was arbitrary, for an entire matrix:

$$\frac{\partial|\mathbf{Y}|}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{Q}_c \mathbf{Y})}{\partial \mathbf{X}} \tag{21}$$

Definition

Let's say that we have

$$\mathbf{A} = \mathbf{X} + \mathbf{E}$$

where $\mathbf{A}$ is the observation matrix, $\mathbf{E}$ is a matrix of stochastic fluctuations with a mean of 0, and $\mathbf{X}$ is our approximation to $\mathbf{A}$.

In Least Squares, our objective is to minimize the sum of squared errors:

$$\begin{aligned} SSE &= e_{11}^2 + e_{12}^2 + \cdots + e_{1n}^2 + e_{21}^2 + \cdots + e_{2n}^2 + \cdots + e_{mn}^2 \\ &= \sum_i \sum_j e_{ij}^2 \\ &= \sum_i \sum_j (e_{ij} e_{ij}) \\ &= \mathrm{tr}(\mathbf{E}^T \mathbf{E}) && \text{by (5)} \end{aligned}$$

Definition

If we have no constraints on $\mathbf{X}$, then we are, equivalently, minimizing:

$$SSE = \text{tr}(\mathbf{E}^T\mathbf{E}) = \text{tr}[(\mathbf{A} - \mathbf{X})^T(\mathbf{A} - \mathbf{X})]$$

Calculation

A minimization:

$$\begin{aligned}
\frac{\partial(SSE)}{\partial \mathbf{X}} &= \frac{\partial \operatorname{tr}(\mathbf{E}^T\mathbf{E})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}[(\mathbf{A}-\mathbf{X})^T(\mathbf{A}-\mathbf{X})]}{\partial \mathbf{X}} \\
&= \frac{\partial \operatorname{tr}(\mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{X} - \mathbf{X}^T\mathbf{A} + \mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} \\
&= \frac{\partial[\operatorname{tr}(\mathbf{A}^T\mathbf{A}) - \operatorname{tr}(\mathbf{A}^T\mathbf{X}) - \operatorname{tr}(\mathbf{X}^T\mathbf{X}) + \operatorname{tr}(\mathbf{X}^T\mathbf{X})]}{\partial \mathbf{X}} \qquad \text{by (1)} \\
&= \frac{\partial \operatorname{tr}(\mathbf{A}^T\mathbf{A})}{\partial \mathbf{X}} - \frac{\partial \operatorname{tr}(\mathbf{A}^T\mathbf{X})}{\partial \mathbf{X}} - \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{A})}{\partial \mathbf{X}} + \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} \\
&= 0 - \frac{\partial \operatorname{tr}(\mathbf{A}^T\mathbf{X})}{\partial \mathbf{X}} - \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{A})}{\partial \mathbf{X}} + \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}}
\end{aligned}$$

Calculation

Continuing:

$$
\begin{aligned}
\frac{\partial(SSE)}{\partial \mathbf{X}} &= -\frac{\partial \operatorname{tr}(\mathbf{A}^T\mathbf{X})}{\partial \mathbf{X}} - \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{A})}{\partial \mathbf{X}} + \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} \\
&= -\mathbf{A} - \frac{\partial \operatorname{tr}(\mathbf{A}^T\mathbf{X})}{\partial \mathbf{X}} + \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} && \text{by (13) \& (3)} \\
&= -\mathbf{A} - \mathbf{A} + \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} && \text{by (13)} \\
&= -\mathbf{A} - \mathbf{A} + \frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{I}\mathbf{X}\mathbf{I})}{\partial \mathbf{X}} \\
&= -\mathbf{A} - \mathbf{A} + [\mathbf{I}^T\mathbf{X}\mathbf{I}^T + \mathbf{I}\mathbf{X}\mathbf{I}] && \text{by (14)} \\
&= -\mathbf{A} - \mathbf{A} + [\mathbf{X} + \mathbf{X}] = -2\mathbf{A} + 2\mathbf{X} && (22)
\end{aligned}
$$

Result

As in any **Least** Squares problem, we should set our derivative equal to 0 in order to find the minimum of the function.

$$\frac{\partial(SSE)}{\partial \mathbf{X}} = -2\mathbf{A} + 2\mathbf{X} = 0$$
$$2\mathbf{X} = 2\mathbf{A}$$
$$\widehat{\mathbf{A}} = \mathbf{X} = \mathbf{A}$$

Surprisingly, without **any** constraints on $\mathbf{X}$, the best approximation of $\mathbf{A}$ is $\mathbf{A}$ itself.

Oh, the things you learn in calculus $\ddot\smile$!

LaGrange Multipliers

Pretend you have a function:

$$f(\mathbf{X})$$

To maximize or minimize less than mn restraints equivalent to

$$h(x_{11}, \ldots, x_{mn})_{ij} = 0$$

use LaGrange Multipliers u_{ij} (one for **each** restraint), and set

$$g(\mathbf{X}) = f(\mathbf{X}) + \sum_i \sum_j (u_{ij} h_{ij}) \tag{23}$$

Finally, take the derivative with respect to $\mathbf{X}$, set equal to 0, and solve.

Definition

We still want to find an $\mathbf{X}$ that minimizes the SSE to best approximate $\mathbf{A}$; however, we are now subject to the constraint that $\mathbf{X}$ is a Symmetric Matrix.

$\mathbf{X}$ Symmetric Means:

$$\mathbf{X} = \mathbf{X}^T$$
$$\mathbf{X} - \mathbf{X}^T = 0$$

The Recipe:

1. We have our equation to minimize: $\text{tr}(\mathbf{E}^T\mathbf{E})$.
2. We have our constraint: $\mathbf{H} = \mathbf{X} - \mathbf{X}^T = 0$.
3. Put in LaGrange multiplier form.
4. Take the derivative.
5. Set the derivative equal to 0.
6. Solve for $\mathbf{X}$.

Calculation

First -- set up the problem:

$$
\begin{aligned}
g(\mathbf{X}) &= f(\mathbf{X}) + \sum_i \sum_j (u_{ij} h_{ij}) && \text{by (23)} \\
&= \mathrm{tr}(\mathbf{E}^T\mathbf{E}) + \mathrm{tr}(\mathbf{U}^T\mathbf{H}) && \text{by (5)} \\
&= \mathrm{tr}(\mathbf{E}^T\mathbf{E}) + \mathrm{tr}[\mathbf{U}^T(\mathbf{X} - \mathbf{X}^T)] \\
&= \mathrm{tr}(\mathbf{E}^T\mathbf{E}) + \mathrm{tr}(\mathbf{U}^T\mathbf{X}) - \mathrm{tr}(\mathbf{U}^T\mathbf{X}^T) && \text{by (1)} \\
&= \mathrm{tr}(\mathbf{E}^T\mathbf{E}) + \mathrm{tr}(\mathbf{U}^T\mathbf{X}) - \mathrm{tr}(\mathbf{U}\mathbf{X}) && \text{by (3) \& (4)}
\end{aligned}
$$

Calculation

Second -- take the derivative:

$$
\begin{aligned}
\frac{\partial g(\mathbf{X})}{\partial \mathbf{X}} &= \frac{\partial[\mathrm{tr}(\mathbf{E}^T\mathbf{E}) + \mathrm{tr}(\mathbf{U}^T\mathbf{X}) - \mathrm{tr}(\mathbf{U}\mathbf{X})]}{\partial \mathbf{X}} \\
&= \frac{\partial\,\mathrm{tr}(\mathbf{E}^T\mathbf{E})}{\partial \mathbf{X}} + \frac{\partial\,\mathrm{tr}(\mathbf{U}^T\mathbf{X})}{\partial \mathbf{X}} - \frac{\partial\,\mathrm{tr}(\mathbf{U}\mathbf{X})}{\partial \mathbf{X}} \\
&= -2\mathbf{A} + 2\mathbf{X} + \frac{\partial\,\mathrm{tr}(\mathbf{U}^T\mathbf{X})}{\partial \mathbf{X}} - \frac{\partial\,\mathrm{tr}(\mathbf{U}\mathbf{X})}{\partial \mathbf{X}} && \text{by (22)} \\
&= -2\mathbf{A} + 2\mathbf{X} + \mathbf{U} - \mathbf{U}^T && \text{by (13)}
\end{aligned}
$$

Calculation

Third -- set the derivative equal to 0:

$$\frac{\partial g(\mathbf{X})}{\partial \mathbf{X}} = -2\mathbf{A} + 2\mathbf{X} + \mathbf{U} - \mathbf{U}^T = 0$$

$$2\mathbf{X} = 2\mathbf{A} + \mathbf{U}^T - \mathbf{U}$$

$$\mathbf{X} = \mathbf{A} + \frac{\mathbf{U}^T - \mathbf{U}}{2}$$

However, now note that: $\mathbf{X} = \mathbf{X}^T$

$$\mathbf{X}^T = \left(\mathbf{A} + \frac{\mathbf{U}^T - \mathbf{U}}{2}\right)^T$$

$$\mathbf{X}^T = \mathbf{A}^T + \frac{\mathbf{U} - \mathbf{U}^T}{2}$$

Result

Fourth -- add $\mathbf{X}^T$ to both sides and solve for $\mathbf{X}$:

$$\begin{aligned}
\mathbf{X} + \mathbf{X}^T &= \mathbf{A} + \frac{\mathbf{U}^T - \mathbf{U}}{2} + \mathbf{A}^T + \frac{\mathbf{U} - \mathbf{U}^T}{2} \\
\mathbf{X} + \mathbf{X} &= \mathbf{A} + \mathbf{A}^T + \frac{\mathbf{U}^T - \mathbf{U}}{2} - \frac{\mathbf{U}^T - \mathbf{U}}{2} \\
2\mathbf{X} &= \mathbf{A} + \mathbf{A}^T \\
\widehat{\mathbf{A}} = \mathbf{X} &= \frac{\mathbf{A} + \mathbf{A}^T}{2}
\end{aligned}$$

Therefore, to approximate $\mathbf{A}$ with a Symmetric Matrix, the **best** matrix (according to the Least Squares Criterion) is the **average** of the elements of $\mathbf{A}$ **and** the elements of $\mathbf{A}^T$.

Definition

The function we want to maximize:

$$u = |\mathbf{U}^{-1}(\mathbf{R} - \mathbf{F}\mathbf{F}^T)\mathbf{U}^{-1}|$$

1. $\mathbf{R}$ is a correlation matrix.
   - $\text{diag}(\mathbf{R}) = \mathbf{I}$
2. $\mathbf{F}$ is a factor pattern matrix of **uncorrelated** common factors.
3. $\mathbf{U}^2$ is a covariance matrix of **uncorrelated** unique factors.
   - $\mathbf{U}^2 = \text{diag}(\mathbf{U}^2) = \mathbf{I} - \text{diag}(\mathbf{F}\mathbf{F}^T)$

Definition

The function we want to maximize:

$$u = |\mathbf{U}^{-1}(\mathbf{R} - \mathbf{F}\mathbf{F}^T)\mathbf{U}^{-1}|$$

The function u is a likelihood ratio criterion for a test of independence after the common factors have been partialed out of the covariance matrix.

$\mathbf{U}^{-1}(\mathbf{R} - \mathbf{F}\mathbf{F}^T)\mathbf{U}^{-1}$ should be close to $\mathbf{I}$, so u should be close to 1.

We want to find the $\mathbf{F}$ (and consequently the $\mathbf{U}^2$) that results in a determinant as close to 1 as possible.

Definition

To make the derivatives simpler, let:

$$u_1 = |\mathbf{U}^{-2}| \qquad \text{and} \qquad u_2 = |\mathbf{R} - \mathbf{F}\mathbf{F}^T|$$

Note that:

$$u_1 u_2 = |\mathbf{U}^{-2}||\mathbf{R} - \mathbf{F}\mathbf{F}^T| = |\mathbf{U}^{-1}||\mathbf{R} - \mathbf{F}\mathbf{F}^T||\mathbf{U}^{-1}| = |\mathbf{U}^{-1}(\mathbf{R} - \mathbf{F}\mathbf{F}^T)\mathbf{U}^{-1}|$$

Thus, we can use the product rule to find the derivative:

$$\begin{aligned}\frac{\partial u}{\partial \mathbf{F}} &= \frac{\partial (u_1 u_2)}{\partial \mathbf{F}} \\ &= \frac{\partial u_1}{\partial \mathbf{F}} u_2 + u_1 \frac{\partial u_2}{\partial \mathbf{F}} && \text{by (7)}\end{aligned}$$

Calculation: $\frac{\partial u_1}{\partial \mathbf{F}}$

Let's find the derivative of the first part.

$$\begin{aligned}
\frac{\partial u_1}{\partial \mathbf{F}} = \frac{\partial|\mathbf{U}^{-2}|}{\partial \mathbf{F}} &= \frac{\partial|\mathbf{U}^2|^{-1}}{\partial \mathbf{F}} && \text{(by determinant rules)} \\
&= \frac{\partial \operatorname{tr}(|\mathbf{U}^2|^{-1})}{\partial \mathbf{F}} && \text{(since } \operatorname{tr}(|\mathbf{X}|) = |\mathbf{X}|) \\
&= -\frac{\operatorname{tr}(|\mathbf{U}^2|_c^{-2}|\mathbf{U}^2|)}{\partial \mathbf{F}} && \text{by (16)} \\
&= -|\mathbf{U}^2|^{-2}\frac{\partial|\mathbf{U}|^2}{\partial \mathbf{F}} && (24)
\end{aligned}$$

Our next objective is to find the derivative of the highlighted part.

Calculation: $\frac{\partial u_1}{\partial \mathbf{F}}$

Continuing:

$$\begin{aligned}
\frac{\partial |\mathbf{U}|}{\partial \mathbf{F}} &= \frac{\partial |\mathbf{I} - \operatorname{diag}(\mathbf{F}\mathbf{F}^T)|}{\partial \mathbf{F}} && \text{(by definition)} \\
&= \frac{\partial \operatorname{tr}\left(\mathbf{Q}_c[\mathbf{I} - \operatorname{diag}(\mathbf{F}\mathbf{F}^T)]\right)}{\partial \mathbf{F}} && \text{by (21)} \\
&= \frac{\partial \operatorname{tr}(\mathbf{Q}_c)}{\partial \mathbf{F}} - \frac{\partial \operatorname{tr}[\mathbf{Q}_c \operatorname{diag}(\mathbf{F}\mathbf{F}^T)]}{\partial \mathbf{F}} && \text{by (1) \& (2)} \\
&= 0 - \frac{\partial \operatorname{tr}[\mathbf{Q}_c \operatorname{diag}(\mathbf{F}\mathbf{F}^T)]}{\partial \mathbf{F}} \\
&= -\frac{\partial \operatorname{tr}[\mathbf{Q}_c \operatorname{diag}(\mathbf{F}\mathbf{F}^T)]}{\partial \mathbf{F}}
\end{aligned}$$

Calculation: $\frac{\partial u_1}{\partial \mathbf{F}}$

Based on (21), $\mathbf{Q}_c$ is the Adjoint of $[\mathbf{I} - \text{diag}(\mathbf{F}\mathbf{F}^T)] = \mathbf{U}^2$.

Therefore:

$$|\mathbf{U}^2|^{-1}\mathbf{Q}_c = (\mathbf{U}^2)^{-1} \qquad \text{by (17)}$$
$$\mathbf{Q}_c = |\mathbf{U}^2|(\mathbf{U}^{-2})$$

Because $\mathbf{U}^2$ is a diagonal matrix, $\mathbf{Q}_c = |\mathbf{U}^2|(\mathbf{U}^{-2})$ is a diagonal matrix.

And:

$$\frac{\partial |\mathbf{U}|}{\partial \mathbf{F}} = -\frac{\partial \operatorname{tr}[\mathbf{Q}_c \operatorname{diag}(\mathbf{F}\mathbf{F}^T)]}{\partial \mathbf{F}} = -\frac{\partial \operatorname{tr}[\operatorname{diag}(\mathbf{Q}_c\mathbf{F}\mathbf{F}^T)]}{\partial \mathbf{F}} = -\frac{\partial \operatorname{tr}(\mathbf{Q}_c\mathbf{F}\mathbf{F}^T)}{\partial \mathbf{F}}$$

Because the trace only **operates** on **the diagonal**, the trace of the diagonal of a matrix is the same as the trace of the original matrix.

Calculation: $\frac{\partial u_1}{\partial \mathbf{F}}$

Continuing:

$$
\begin{aligned}
-\frac{\partial \operatorname{tr}(\mathbf{Q}_c\mathbf{F}\mathbf{F}^T)}{\partial \mathbf{F}} &= -\frac{\partial \operatorname{tr}(\mathbf{F}^T\mathbf{Q}_c\mathbf{F}\mathbf{I})}{\partial \mathbf{F}} && \text{by (4)} \\
&= -(\mathbf{Q}^T\mathbf{F}\mathbf{I}^T + \mathbf{Q}\mathbf{F}\mathbf{I}) && \text{by (14)} \\
&= -(\mathbf{Q}^T + \mathbf{Q})\mathbf{F} \\
&= -2\mathbf{Q}\mathbf{F} && (\mathbf{Q} \text{ is symmetric}) \\
&= -2|\mathbf{U}^2|\mathbf{U}^{-2}\mathbf{F} && \text{by (17)}
\end{aligned}
$$

And, thus:

$$
\begin{aligned}
\frac{\partial u_1}{\partial \mathbf{F}} &= -|\mathbf{U}^2|\frac{\partial |\mathbf{U}|^2}{\partial \mathbf{F}} && \text{by (24)} \\
&= -|\mathbf{U}^2|^{-2}(-2|\mathbf{U}^2|\mathbf{U}^{-2}\mathbf{F}) \\
&= 2|\mathbf{U}^2|^{-1}\mathbf{U}^{-2}\mathbf{F} \\
&= 2|\mathbf{U}^{-2}|\mathbf{U}^{-2}\mathbf{F} && (25)
\end{aligned}
$$

Calculation: $\frac{\partial u_2}{\partial \mathbf{F}}$

Now, let's find the derivative of the second part.

$$\begin{aligned}
\frac{\partial u_2}{\partial \mathbf{F}} &= \frac{\partial |\mathbf{R} - \mathbf{F}\mathbf{F}^T|}{\partial \mathbf{F}} \\
&= \frac{\partial \operatorname{tr}[\mathbf{Q}_c(\mathbf{R} - \mathbf{F}\mathbf{F}^T)]}{\partial \mathbf{F}} && \text{by (21)} \\
&= \frac{\partial \operatorname{tr}(\mathbf{Q}_c\mathbf{R})}{\partial \mathbf{F}} - \frac{\partial \operatorname{tr}(\mathbf{Q}_c\mathbf{F}\mathbf{F}^T)}{\partial \mathbf{F}} && \text{by (1) \& (2)} \\
&= 0 - \frac{\partial \operatorname{tr}(\mathbf{F}^T\mathbf{Q}_c\mathbf{F}\mathbf{I})}{\partial \mathbf{F}} && \text{by (4)} \\
&= -(\mathbf{Q}^T + \mathbf{Q})\mathbf{F} && \text{by (14)} \\
&= -2\mathbf{Q}\mathbf{F} && (\mathbf{Q} \text{ is symmetric}) \\
&= -2|\mathbf{R} - \mathbf{F}\mathbf{F}^T|(\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F} && (26)
\end{aligned}$$

Calculation: Entire Thing

Putting the pieces together:

$$\frac{\partial u}{\partial \mathbf{F}} = \frac{\partial u_1}{\partial \mathbf{F}} u_2 + u_1 \frac{\partial u_2}{\partial \mathbf{F}}$$

Which implies that

$$\frac{\partial u}{\partial \mathbf{F}} = (2|\mathbf{U}^{-2}|\mathbf{U}^{-2}\mathbf{F})|\mathbf{R} - \mathbf{F}\mathbf{F}^T| \\ + |\mathbf{U}^{-2}|(-2|\mathbf{R} - \mathbf{F}\mathbf{F}^T|(\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F})$$

Calculation: Finding Maximum

To find the maximum of this function, we must set it equal to 0 and solve.

$$
\begin{aligned}
0 &= 2|\mathbf{U}^{-2}|(\mathbf{U}^{-2}\mathbf{F})|\mathbf{R} - \mathbf{F}\mathbf{F}^T| - 2|\mathbf{U}^{-2}||\mathbf{R} - \mathbf{F}\mathbf{F}^T|(\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F} \\
0 &= 2|\mathbf{U}^{-2}||\mathbf{R} - \mathbf{F}\mathbf{F}^T|(\mathbf{U}^{-2}\mathbf{F} - (\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F}) \\
0 &= \mathbf{U}^{-2}\mathbf{F} - (\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F}
\end{aligned}
$$

Calculation: Finding Maximum

Finishing the calculation:

$$
\begin{aligned}
0 &= \mathbf{U}^{-2}\mathbf{F} - (\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F} \\
(\mathbf{R} - \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{F} &= \mathbf{U}^{-2}\mathbf{F} \\
\mathbf{F} &= (\mathbf{R} - \mathbf{F}\mathbf{F}^T)\mathbf{U}^{-2}\mathbf{F} \\
\mathbf{F} &= \mathbf{R}\mathbf{U}^{-2}\mathbf{F} - \mathbf{F}\mathbf{F}^T\mathbf{U}^{-2}\mathbf{F} \\
\mathbf{R}\mathbf{U}^{-2}\mathbf{F} - \mathbf{F} &= \mathbf{F}\mathbf{F}^T\mathbf{U}^{-2}\mathbf{F} \\
(\mathbf{R}\mathbf{U}^{-2} - \mathbf{I})\mathbf{F} &= \mathbf{F}(\mathbf{F}^T\mathbf{U}^{-2}\mathbf{F}) \qquad (27)
\end{aligned}
$$

Result

Based on the previous slide, we have

$$(\mathbf{R}\mathbf{U}^{-2} - \mathbf{I})\mathbf{F} = \mathbf{F}(\mathbf{F}^T\mathbf{U}^{-2}\mathbf{F})$$

Let $\mathbf{\Lambda} = (\mathbf{F}^T\mathbf{U}^{-2}\mathbf{F})$ be diagonal. Then

$$(\mathbf{R}\mathbf{U}^{-2} - \mathbf{I})(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n) = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n)\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \vdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

$$(\mathbf{R}\mathbf{U}^{-2} - \mathbf{I})(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n) = (\mathbf{f}_1\lambda_1, \mathbf{f}_2\lambda_2, \dots, \mathbf{f}_n\lambda_n)$$

$$(\mathbf{R}\mathbf{U}^{-2} - \mathbf{I})(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n) = (\lambda_1\mathbf{f}_1, \lambda_2\mathbf{f}_2, \dots, \lambda_n\mathbf{f}_n)$$

is an implicit eigenproblem.

References

- Schönemann, P. H. (1985). On the formal differentiation of traces and determinants. *Multivariate Behavioral Research*, *20*, 113–139.
- Schönemann, P. H. (2001). Better never than late: Peer review and the preservation of prejudice. *Ethical Human Sciences and Services*, *3*, 7–21.