VECTOR / MATRIX DERIVATIVES

Everything You Need To Know

KANG GYEONGHUN

ESC, YONSEI UNIVERSITY

March 29, 2020

Table of Contents

1. Vector Derivatives
   - Directional Derivative
   - Jacobian Matrix
   - Vector Derivatives Rules
   - Example: Least Sqaures Estimators
2. Change of Variables
   - Why do we need Jacobian?
   - Example: Geometric Interpretation of MVN
3. Matrix Derivatives
   - Directional Derivative in Matrix space
   - Matrix Derivatives for Determinant
   - Matrix Derivatives for Trace
   - Example: MLE for MVN

I. Vector Derivatives

DIRECTIONAL DERIVATIVE

Taylor Expansion for Derivatives

- Suppose $f(x) \in C^2[a,b]$. We consider **Taylor expansion** of the function f at w;

$$f(x) \approx f(w) + \frac{(x-w)}{1!} f'(w) + \frac{(x-w)^2}{2!} f''(w)$$

- Then for sufficiently small e, we have;

$$f(x+e) \approx f(w) + \frac{(x+e-w)}{1!} f'(w) + \frac{(x+e-w)^2}{2!} f''(w)$$

- With respect to e, we can say

$$f(x+e) \approx O(1) + O(e) + O(e^2)$$

This formulation provides us an intuitive framework in understanding derivatives in various spaces.

DIRECTIONAL DERIVATIVE

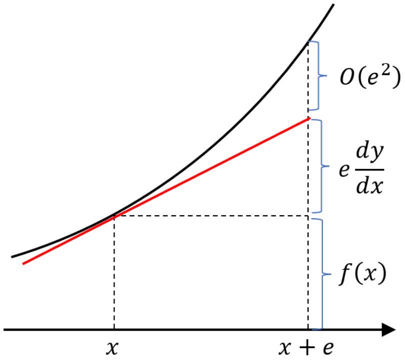
Recall: 1-dimensional Derivative

- For $f(x) : \mathbb{R} \mapsto \mathbb{R}$, an infinitsimal change e in the input x leads to change in the output;

$$f(x+e) = f(x) + e\frac{df}{dx} + O(e^2)$$

- Rearrage, $e \to 0$, and we have 1D Derivative;

$$\frac{dy}{dx} = \lim_{e \to 0} \frac{f(x+e) - f(x)}{e}$$



Scalar field

DIRECTIONAL DERIVATIVE

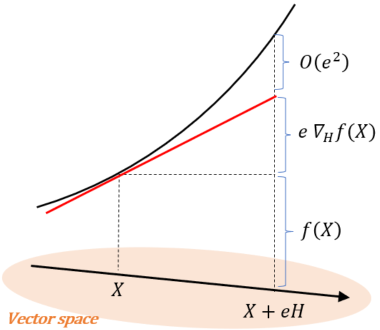
n-dimensional Derivative

- For $f(x) : \mathbb{R}^n \mapsto \mathbb{R}$, an infinitsimal change e in the input $\mathbf{x}$ **in the direction of $\mathbf{H}$** leads to change in the output;

$$f(\mathbf{X} + e\mathbf{H}) = f(\mathbf{X}) + e\nabla_{\mathbf{H}} f(\mathbf{X}) + O(e^2)$$

- Rearrage, $e \to 0$, and we have nD **Directional** Derivative;

$$\nabla_{\mathbf{H}} f(\mathbf{X}) = \lim_{e \to 0} \frac{f(\mathbf{X} + e\mathbf{H}) - f(\mathbf{X})}{e}$$



$$\nabla_H f(X) = H \cdot \nabla f = H \cdot \frac{\partial f}{\partial X}$$

DIRECTIONAL DERIVATIVE

Directional Derivative in Vectorspace

- With a gradient of a function $f; \mathbf{x} \mapsto \mathbb{R}$, once we specify a direction $\mathbf{b}$ (unit vector) of variation in the input $\mathbf{x}$, we have a **directional derivative of** f;

$$\begin{aligned} \nabla_{\mathbf{b}} f(\mathbf{x}) &= f'(\mathbf{x}) \cdot \mathbf{b} = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{b}) - f(\mathbf{x})}{\epsilon} \\ &= \nabla f(\mathbf{x}) \cdot \mathbf{b} = \|\nabla f(\mathbf{x})\| \|\mathbf{b}\| \cos\theta \end{aligned}$$

- Directional derivative is just a $\mathbb{R}^1$ derivative generalized to $\mathbb{R}^n$ space. The difference is that, unlike in $\mathbb{R}^1$ where we had only a number line, in higher dimensions we have to specify which direction $d\mathbf{x}$ is headed.

1D derivatives: $$\frac{df(x)}{dx} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

- $\nabla f(\mathbf{x})$ is called **gradient** and represents a direction in $\mathbf{x}$ of the maximum change in $f(\mathbf{x})$ since $\cos\theta = 1 \iff \theta = 0$.

JACOBIAN MATRIX

Jacobian Matrix for Multidimensional Output

Define $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by $\mathbf{y} = \psi(\mathbf{x})$. The $m \times n$ matrix of first-order derivatives of this transformation is called **Jacobian Matrix** and is expressed as;

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \nabla y_1 \\ \nabla y_2 \\ \vdots \\ \nabla y_m \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}_{m \times n}$$

- If $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^1$, we have a single row $\dfrac{\partial y}{\partial \mathbf{x}} = \nabla \psi(\mathbf{x}) = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$, which is called **Gradient of** y.
- Jacobian Matrix for $m \times 1$ vector if essentially a vertical stack of each element's gradient.

JACOBIAN MATRIX

Jacobian Matrix for Multidimensional Output

Define $\psi : \mathbb{R}^n \to \mathbb{R}^m$ by $\mathbf{y} = \psi(\mathbf{x})$. The $m \times n$ matrix of first-order derivatives of this transformation is called **Jacobian Matrix** and is expressed as;

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}_{m \times n}$$

- If $\psi : \mathbb{R}^1 \to \mathbb{R}^m$, we have a single column $\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$

VECTOR DERIVATIVES

Vector Derivatives Rule 1

Let $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{y} \in \mathbb{R}^{m\times 1}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$. ($\mathbf{A}$ is constant matrix.)

- **pf)** The easiest way is by showing element-wise operations.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

So we have $y_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n$, and $\dfrac{\partial y_i}{\partial x_j} = a_{ij}$.

Therefore, $\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \dots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} = \dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$. $\square$

VECTOR DERIVATIVES

Vector Derivatives Rule 2

Let $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{y} \in \mathbb{R}^{m\times 1}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$ ($\mathbf{x}$ is a function of $\mathbf{z}$), and $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{A}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$. ($\mathbf{A}$ is constant matrix.)

- **pf)** Since $y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n$, we have $\dfrac{\partial y_i}{\partial z_j} = a_{i1}\dfrac{\partial x_1}{\partial z_j} + a_{i2}\dfrac{\partial x_2}{\partial z_j} + \ldots + a_{in}\dfrac{\partial x_n}{\partial z_j}$.

  Since $\dfrac{\partial x_i}{\partial z_j}$ is (i,j)th element of $\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$,

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}}_{(i,j)} = \frac{\partial y_i}{\partial z_j} = \sum_{k=1}^{n} a_{ik}\frac{\partial x_k}{\partial z_j} = \sum_{k=1}^{n} \mathbf{A}_{(i,k)}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}_{(k,j)} = \frac{\partial \mathbf{x}}{\partial \mathbf{z}}_{(i,j)} \qquad \square$$

- This serves as a proof of vector chain rule: $\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}} = \dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{A}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$.

VECTOR DERIVATIVES

Vector Derivatives Rule 3

Let $\alpha = \mathbf{y}^{\mathbf{T}}\mathbf{A}\mathbf{x}$ where $\mathbf{y} \in \mathbb{R}^{m\times 1}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, and $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\dfrac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^{\mathbf{T}}\mathbf{A}$ and $\dfrac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}$

($\mathbf{A}$ is constant matrix.)

- **pf)** Since $\alpha = \mathbf{y}^{\mathbf{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{y}$, we only prove $\dfrac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^{\mathbf{T}}\mathbf{A}$.

$$\alpha = \mathbf{y}^{\mathbf{T}}\mathbf{A}\mathbf{x} = \mathbf{w}^{\mathbf{T}}\mathbf{x} = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

$$\frac{\partial \alpha}{x_j} = w_j = [y^T A]_j$$

$$\therefore \frac{\partial \alpha}{\partial \mathbf{x}_j} = [\mathbf{y}^{\mathbf{T}}\mathbf{A}]_j \implies \frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^{\mathbf{T}}\mathbf{A}$$

$\square$

VECTOR DERIVATIVES

Vector Derivatives Rule 4 (Quadratic formula)

Let $\alpha = \mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^{n\times 1}$, and $\mathbf{A} \in \mathbb{R}^{n\times n}$. Then $\dfrac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^\mathbf{T}(\mathbf{A} + \mathbf{A}^\mathbf{T})$. ($\mathbf{A}$ is constant matrix.)

- **pf)** It helps to see what α is made of.

$$\alpha = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n x_i a_{i1} & \sum_{i=1}^n x_i a_{i2} & \dots & \sum_{i=i}^n x_i a_{in} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{j=1}^n x_j \sum_{i=1}^n x_i a_{ij} = \sum_{j=1}^n \sum_{i=1}^n x_i x_j a_{ij}$$

VECTOR DERIVATIVES

Vector Derivatives Rule 4 (Quadratic formula) (continued)

Let $\alpha = \mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^{n \times 1}$, and $\mathbf{A} \in \mathbb{R}^{n \times n}$. Then $\dfrac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^{\mathbf{T}}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})$. ($\mathbf{A}$ is constant matrix.)

- **pf)** It helps to see what α is made of.

$$\alpha = \sum_{j=1}^{n} \sum_{i=1}^{n} x_i x_j a_{ij}$$

Note that each x_j has coefficients 1) along the row ($\sum_{i=1}^{n} a_{ji} x_i$) and 2) along the column ($\sum_{i=1}^{n} a_{ij} x_i$). Therefore,

$$\frac{\partial \alpha}{\partial x_j} = \sum_{i=1}^{n} a_{ji} x_i + \sum_{i=1}^{n} a_{ij} x_i$$

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^{\mathbf{T}}\mathbf{A} + \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}} = \mathbf{x}^{\mathbf{T}}(\mathbf{A} + \mathbf{A}^{\mathbf{T}}) \quad \square$$

VECTOR DERIVATIVES

Vector Derivatives Rule 5 (Quadratic formula)

Let $\alpha = \mathbf{y}^\mathbf{T}\mathbf{x}$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times 1}$, and each is a function of $\mathbf{z}$. Then $\dfrac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\mathbf{T}\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\mathbf{T}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$.

- **pf)** In element-wise expansion,

$$\alpha = x_1y_1 + x_2y_2 + \ldots + x_ny_n$$

$$\frac{\partial \alpha}{\partial z_k} = \sum_{i=1}^{n} y_i \frac{\partial x_i}{\partial z_k} + \sum_{j=1}^{n} x_i \frac{\partial y_j}{\partial z_k} = \sum_{i=1}^{n} y_i \frac{\partial \mathbf{x}}{\partial \mathbf{z}}_{(i,k)} + \sum_{i=1}^{n} y_i \frac{\partial \mathbf{y}}{\partial \mathbf{z}}_{(j,k)}$$

$$= [\mathbf{y}^\mathbf{T}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}]_k + [\mathbf{x}^\mathbf{T}\frac{\partial \mathbf{y}}{\partial \mathbf{z}}]_k \quad \square$$

Or with the chain rule, $\dfrac{\partial \alpha}{\partial \mathbf{z}} = \dfrac{\partial \alpha}{\partial \mathbf{y}}\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}} + \dfrac{\partial \alpha}{\partial \mathbf{x}}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{x}^\mathbf{T}\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\mathbf{T}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$. $\square$

VECTOR DERIVATIVES

Vector Derivatives Rule 6 (Quadratic formula)

Let $\alpha = \mathbf{y}^\mathbf{T}\mathbf{A}\mathbf{x}$ where $\mathbf{y} \in \mathbb{R}^{m\times 1}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, both a function of $\mathbf{z}$ and $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{y}^\mathbf{T}\mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^\mathbf{T}\mathbf{A}^\mathbf{T}\frac{\partial \mathbf{y}}{\partial \mathbf{z}}$. ($\mathbf{A}$ is constant matrix.)

- **pf)** Let $\mathbf{y}^\mathbf{T}\mathbf{A} = \mathbf{w}^\mathbf{T}$. By the last rule,

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{w}^\mathbf{T}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^\mathbf{T}\frac{\partial \mathbf{w}}{\partial \mathbf{z}} = \mathbf{y}^\mathbf{T}\mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^\mathbf{T}\frac{\partial \mathbf{A}^\mathbf{T}\mathbf{y}}{\partial \mathbf{z}} = \mathbf{y}^\mathbf{T}\mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^\mathbf{T}\mathbf{A}^\mathbf{T}\frac{\partial \mathbf{y}}{\partial \mathbf{z}} \quad \square$$

Or with the chain rule,

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \frac{\partial \alpha}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \frac{\partial \alpha}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{y}^\mathbf{T}\mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^\mathbf{T}\mathbf{A}^\mathbf{T}\frac{\partial \mathbf{y}}{\partial \mathbf{z}}. \quad \square$$

VECTOR DERIVATIVES

Vector Derivatives Rule 7 (Quadratic formula)

Let $\alpha = \mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^{n\times 1}$, a function of $\mathbf{z}$, and $\mathbf{A} \in \mathbb{R}^{n\times n}$. Then $\dfrac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^{T}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$.
($\mathbf{A}$ is constant matrix.)

- **pf)** Since if $\alpha = \mathbf{y}^{\mathbf{T}}\mathbf{A}\mathbf{x}$, then $\dfrac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{y}^{\mathbf{T}}\mathbf{A}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}}$, we have

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^{\mathbf{T}}\mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{x}^{T}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})\frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad \square$$

LEAST SQUARES ESTIMATORS

Example: Least Sqaures Estimators

- For a target vector $\mathbf{t}$ and a design matrix $\mathbf{\Phi}$, LSE $\mathbf{w}$ can be obatained by;

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 = (\mathbf{t} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w})$$

  Differentiate wrt $\mathbf{w}$ and we have

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{t}^{\mathbf{T}}\mathbf{t} - 2\mathbf{t}^{\mathbf{T}}\mathbf{\Phi}\mathbf{w} + \mathbf{w}^{\mathbf{T}}\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi}\mathbf{w}) =^{set} 0$$

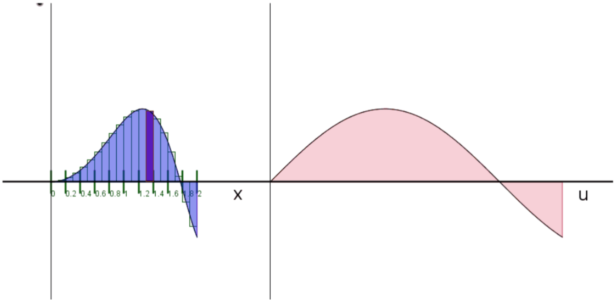
$$\therefore \mathbf{w}_{OLS} = (\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathbf{T}}\mathbf{t}$$

II. Change of Variables

WHY NEED JACOBIAN?

Recall: Change of Variable Technique for Integration

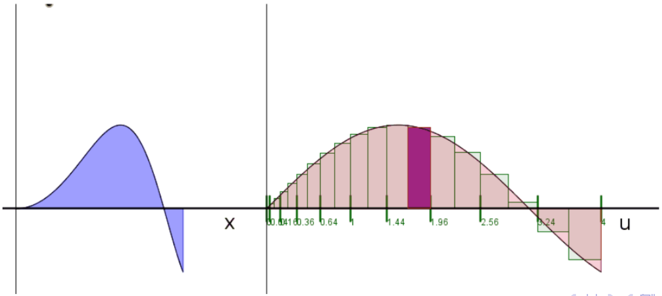
- For $\int_0^2 \sin(x^2)dx$, substitute $u := x^2$ to have $\int_0^2 \sin(x^2)dx = \int_0^4 \sin(u)dx$.
- By substituting $u := x^2$, the coordinate space has been extended by $\dfrac{du}{dx}$.



WHY NEED JACOBIAN?

Recall: Change of Variable Technique for Integration

- We multiply by $\frac{dx}{du}$ to mitigate this elongation; $\int_0^4 \sin(u)dx = \int_0^4 \sin(u)\frac{dx}{du}du = \int_0^4 \frac{\sin(u)}{2\sqrt{u}}du$
- Generalization of $\frac{dx}{du}$ to nd space is det of Jacobian; $|\frac{\partial \mathbf{x}}{\partial \mathbf{u}}|$. **It comes from the amount that the area is stretched under the coordinate transformation.** (source)



WHY NEED JACOBIAN?

Change of Variable in Probability Distribution

- COV for probability distribution is identical to COV for definite integral. For $x \sim p(x)$, $u := f(x) \sim p(f^{-1}(u))|\frac{dx}{du}|$.
- **Example:** Let $x \sim 3x^2$ over $[0, 1]$. Define $u := x^2$. Then $u \sim 3u|\frac{1}{2}y^{-1/2}| = \frac{3}{2}u^{1/2}$.
- For multivariable case where $\mathbf{y} = \mathbf{f}(\mathbf{x})$, we have

$$p(\mathbf{y}) = p(\mathbf{x})|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}|$$

where $p(\mathbf{x})$ is expressed in terms of $\mathbf{y}$.

MVN: GEOMETRIC INTERPRETATION

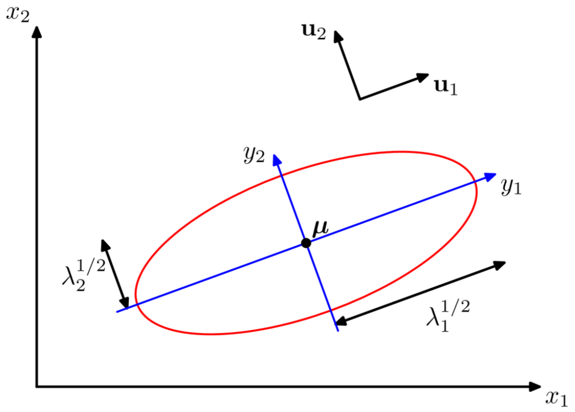
Change of Basis to Eigenvectors of Σ

$$\textbf{MVN:} \quad N(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{(\Sigma)^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{x}-\mu)\}$$

- Since $\mathbf{\Sigma}$ is symmetric and positive definite, we have eigenvalue decomposition of $\mathbf{\Sigma}$ with strictly positive λ_i; $\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, and $\mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$
- Substitute $\mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$ and we have $(\mathbf{x}-\mu)^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{x}-\mu) = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$, where we defined $y_i := \mathbf{u}_i^T(\mathbf{x}-\mu)$, which is a coordinate of $(\mathbf{x}-\mu)$ projected to a unit vector $\mathbf{u}$.
- Define $\mathbf{y} := \mathbf{U}^T(\mathbf{x}-\mu)$ where $\mathbf{U}$ has $\mathbf{u}_i$ as column. Then we have a det of jocobian of 1 since $|\mathbf{J}|^2 = |\mathbf{U}^{\mathbf{T}}|^2 = |\mathbf{U}^{\mathbf{T}}||\mathbf{U}| = |\mathbf{I}|$. Thus we have $p(\mathbf{y}) = \prod_{i=1}^{D} \frac{1}{(2\pi\lambda_i)^{1/2}} \exp\{-\sum_{i=1}^{D} \frac{y_i^2}{2\lambda_i}\}$.

MVN: GEOMETRIC INTERPRETATION

Change of Basis to Eigenvectors of Σ



PCA derives from the exactly same logic, except that sample cov matrix $\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi}$ is used in lieu of Σ.

III. Matrix Derivatives

MATRIX DERIVATIVES

Matrix Derivatives

Let $\mathbf{A}$ be $m \times n$ matrix. Derivative of $\mathbf{A}$ with respect to a scalar k is;

$$\frac{\partial \mathbf{A}}{\partial \alpha} = \begin{bmatrix} \frac{\partial a_{11}}{\partial \alpha} & \frac{\partial a_{12}}{\partial \alpha} & \cdots & \frac{\partial a_{1n}}{\partial \alpha} \\ \frac{\partial a_{21}}{\partial \alpha} & \frac{\partial a_{22}}{\partial \alpha} & \cdots & \frac{\partial a_{2n}}{\partial \alpha} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{m1}}{\partial \alpha} & \frac{\partial a_{m2}}{\partial \alpha} & \cdots & \frac{\partial a_{mn}}{\partial \alpha} \end{bmatrix}_{m \times n}$$

Matrix Derivatives Rule:

Let $\mathbf{A}$ be nonsingular. Then $\dfrac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1}\dfrac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1}$

- **pf)** Since $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, $\mathbf{A}^{-1}\dfrac{\partial \mathbf{A}}{\partial \alpha} + \dfrac{\partial \mathbf{A}^{-1}}{\partial \alpha}\mathbf{A} = \mathbf{0}$. Rearrange and we have $\dfrac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1}\dfrac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1}$.
  $\square$

MATRIX DERIVATIVES: DETERMINANT

Directional Derivative in Matrix Space

- Since $\det : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$, we first need to define directional derivative in matrix space. For a mapping $f : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ from a matrix to a real number, a differential in f in response to an infinitesimal change in the input $\mathbf{X}$ in the direction of $\mathbf{H}$ is defined as;

  $$\text{Directional Derivative in Matrix space:} \quad \nabla_{\mathbf{H}} f(\mathbf{X}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{X} + \epsilon\mathbf{H}) - f(\mathbf{X})}{\epsilon}$$

  There are at least dozens of alternative notations for this; $Df(\mathbf{X})(\mathbf{H})$, $f'(\mathbf{X})\mathbf{H}$ ($f'(\mathbf{X})\mathbf{H}$ is NOT a matrix multiplication. In fact, it should be written $\langle f'(\mathbf{X}), \mathbf{H} \rangle_F$ where $_F$ denotes Frobenius inner product.)
- For comparison, in scalar space and vector space we had;

  $$\text{Derivative in scalar space:} \quad f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

  $$\text{Derivative in Vector space:} \quad \nabla_{\mathbf{h}} f(\mathbf{X}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon\mathbf{h}) - f(\mathbf{x})}{\epsilon}$$

MATRIX DERIVATIVES: DETERMINANT

Directional Derivative for det

- Armed with this concept, we can define a directional derivative for det;

$$\nabla_{\mathbf{H}} det(\mathbf{X}) = det'(\mathbf{X})\mathbf{H} = \lim_{\epsilon \to 0} \frac{det(\mathbf{X} + \epsilon \mathbf{H}) - det(\mathbf{X})}{\epsilon}$$

  and prove some useful facts.

Fact 1. $det'(\mathbf{I})\mathbf{H} = tr(\mathbf{H})$

- **pf)** $det(\mathbf{I} + \epsilon \mathbf{H}) = 1 + \epsilon\, tr(\mathbf{H}) + o(\epsilon)$, where $o(\epsilon)$ consists of terms with $2 \leq$ degree in ϵ, so when divided by ϵ, gets annihilated as $\epsilon \to 0$. To see this, recall that det is a sum of signed products of one element of each column that are NOT on the same rows. This makes the diagonal product the only term with ϵ of 1st degree. $\square$
- This tells us that tr is more than a boring sum of diagonal element. If $\mathbf{T}$ moves in the direction of $\mathbf{H}$, then its determinant changes accordingly by a factor of $tr(\mathbf{H})$.

MATRIX DERIVATIVES: DETERMINANT

Fact 2. $det'(\mathbf{X})\mathbf{H} = \det(\mathbf{X})tr(\mathbf{X}^{-1}\mathbf{H})$

- **pf)** The proof is similar to that of Fact 1;

$$\begin{aligned} det(\mathbf{X} + \epsilon\mathbf{H}) &= det(\mathbf{X})\ det(\mathbf{I} + \epsilon\mathbf{X}^{-1}\mathbf{H}) \\ &= det(\mathbf{X})\ (1 + \epsilon\ tr(\mathbf{X}^{-1}\mathbf{H}) + o(\epsilon)) \\ &= det(\mathbf{X}) + \epsilon\ det(\mathbf{X})\ tr(\mathbf{X}^{-1}\mathbf{H}) + o(\epsilon) \quad \square \end{aligned}$$

Fact 3. $\dfrac{\partial}{\partial \mathbf{X}} \det(\mathbf{X}) = \det(\mathbf{X})\mathbf{X}^{-T}$. Equivalently, $\dfrac{\partial}{\partial \mathbf{X}} \ln\det(\mathbf{X}) = \mathbf{X}^{-T}$.

- **pf)** We use $\langle \mathbf{A}, \mathbf{B} \rangle_F = tr(\mathbf{A}^T\mathbf{B})$. From the Fact 2,

$$\begin{aligned} det'(\mathbf{X})\mathbf{H} = \langle det'(\mathbf{X}), \mathbf{H} \rangle_F &= \det(\mathbf{X})tr(\mathbf{X}^{-1}\mathbf{H}) \\ &= \langle \det(\mathbf{X})\mathbf{X}^{-T}, \mathbf{H} \rangle_F \quad \square \end{aligned}$$

MATRIX DERIVATIVES: DETERMINANT

Fact 4. $\frac{\partial}{\partial x} det(\mathbf{X}) = \det(\mathbf{X}) tr(\mathbf{X}^{-1} \frac{\partial \mathbf{X}}{\partial x})$.

- **pf)** This follows from Fact 2. We had;

$$det'(\mathbf{X})\mathbf{H} = \det(\mathbf{X}) tr(\mathbf{X}^{-1}\mathbf{H})$$

Now put $\mathbf{H} = \frac{\partial \mathbf{X}}{\partial x}$, which is a matrix itself, and by the Chain Rule we have;

$$\frac{\partial det \mathbf{X}}{\partial \mathbf{X}} \frac{\partial \mathbf{X}}{\partial x} = \frac{\partial}{\partial x} det(\mathbf{X}) = \det(\mathbf{X}) tr(\mathbf{X}^{-1} \frac{\partial \mathbf{X}}{\partial x}) \quad \square$$

MATRIX DERIVATIVES: TRACE

These are relatively intuitive, and the proof is literally no more than laying out elementwise multiplications.

$$\frac{\partial}{\partial A_{ij}} tr(\mathbf{AB}) = B_{ji}$$

$$\frac{\partial}{\partial \mathbf{A}} tr(\mathbf{AB}) = \mathbf{B^T}$$

$$\frac{\partial}{\partial \mathbf{A}} tr(\mathbf{AB^T}) = \frac{\partial}{\partial \mathbf{A}} tr(\mathbf{A^T B}) = \mathbf{B}$$

$$\frac{\partial}{\partial \mathbf{A}} tr(\mathbf{A}) = \mathbf{I}$$

$$\frac{\partial}{\partial \mathbf{A}} tr(\mathbf{ABA^T}) = \mathbf{A}(\mathbf{B} + \mathbf{B^T})$$

MLE FOR MULTIVARIATE GAUSSIAN

Let's say we have a data $\mathbf{X}$ consists of N observations with D features, with each observation is iid sample of $MVN(\mu, \mathbf{\Sigma})$. Then the joint likelihood of $\mathbf{X}$ is given as;

$$p(\mathbf{X}|\mu, \mathbf{\Sigma}) = (2\pi)^{-ND/2}\mathbf{\Sigma}^{-N/2}\prod_{n=1}^{N}\exp(-\frac{1}{2}(\mathbf{x_n} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x_n} - \mu))$$

Taking a log we have

$$\ln p(\mathbf{X}|\mu, \mathbf{\Sigma}) = -\frac{ND}{2}\ln 2\pi - \frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x_n} - \mu)$$

We can get MLE for $\mu, \mathbf{\Sigma}$ by differentiating the above formula by each and find $\mu_{ml}, \mathbf{\Sigma}_{ml}$ that sets it to zero.

MLE FOR MULTIVARIATE GAUSSIAN

Differentiating w.r.t. μ: By the Vector Derivatives Rule 4,

$$\frac{\partial \ln p}{\partial \mu} = \sum_{n=1}^{N} \mathbf{\Sigma}^{-1}(\mathbf{x_n} - \mu) =^{set} 0 \quad \rightarrow \quad \mu_{ml} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$$

Differentiating w.r.t. $\mathbf{\Sigma}$: This is the tricky part. We have;

$$\frac{\partial \ln p}{\partial \mathbf{\Sigma}} = -\frac{N}{2}\frac{\partial \ln |\mathbf{\Sigma}|}{\partial \mathbf{\Sigma}} - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{n=1}^{N}(\mathbf{x_n} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x_n} - \mu)$$

We know from Fact 3 that $\frac{\partial \ln |\mathbf{\Sigma}|}{\partial \mathbf{\Sigma}} = \mathbf{\Sigma}^{-T}$. The second term is a headache. It helps us to know that for any square matrix $\mathbf{A}$ and vector $\mathbf{x}$,

$$\mathbf{x^T A x} = tr(\mathbf{A x x^T})$$

MLE FOR MULTIVARIATE GAUSSIAN

Differentiating w.r.t. $\mathbf{\Sigma}$: (continued)

$$
\begin{aligned}
tr(\mathbf{A}\mathbf{x}\mathbf{x}^{\mathbf{T}}) &= tr(\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1x_1 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & x_2x_2 & \dots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \dots & x_nx_n \end{bmatrix}) \\
&= \sum_{i=1}^{n} x_1 a_{1i} x_i + \dots + \sum_{i=1}^{n} x_n a_{ni} x_i \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} x_j a_{ji} x_i \\
&= \mathbf{x}^{\mathbf{T}} \mathbf{A} \mathbf{x}
\end{aligned}
$$

MLE FOR MULTIVARIATE GAUSSIAN

Differentiating w.r.t. $\mathbf{\Sigma}$: (continued)
Hence we have $(\mathbf{x_n}-\mu)^{\mathbf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x_n}-\mu) = tr(\mathbf{\Sigma}^{-1}(\mathbf{x_n}-\mu)(\mathbf{x_n}-\mu)^T)$, and the second term can be expressed as; $\sum_{n=1}^{N}(\mathbf{x_n}-\mu)^{\mathbf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x_n}-\mu) = Ntr(\mathbf{\Sigma^{-1}S})$, where $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n}-\mu)(\mathbf{x_n}-\mu)^T$.

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{\Sigma_{ij}}}\sum_{n=1}^{N}(\mathbf{x_n}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x_n}-\mu) = N\frac{\partial}{\partial \mathbf{\Sigma_{ij}}}tr(\mathbf{\Sigma^{-1}S}) &= Ntr(\frac{\partial}{\partial \mathbf{\Sigma_{ij}}}\mathbf{\Sigma^{-1}S}) \\
&= -Ntr(\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \mathbf{\Sigma_{ij}}}\mathbf{\Sigma^{-1}S}) \\
&= -Ntr(\frac{\partial \mathbf{\Sigma}}{\partial \mathbf{\Sigma_{ij}}}\mathbf{\Sigma}^{-1}\mathbf{S\Sigma^{-1}}) \\
&= -N(\mathbf{\Sigma^{-1}S\Sigma^{-1}})_{ij}
\end{aligned}
$$

The last line follows since $[\frac{\partial \mathbf{\Sigma}}{\partial \mathbf{\Sigma_{ij}}}]_{ij} = 0$ for all but the position (i, j).

MLE FOR MULTIVARIATE GAUSSIAN

Differentiating w.r.t. Σ: (continued)
This makes the derivative in the second term

$$-\frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{n=1}^{N}(\mathbf{x_n}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x_n}-\mu)=\frac{1}{2}N(\mathbf{\Sigma^{-1}S\Sigma^{-1}})$$

To sum up, rewriting $\frac{\partial \ln p}{\partial \mathbf{\Sigma}}=-\frac{N}{2}\frac{\partial \ln|\mathbf{\Sigma}|}{\partial \mathbf{\Sigma}}-\frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{n=1}^{N}(\mathbf{x_n}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x_n}-\mu)$, we have

$$-\frac{N}{2}\mathbf{\Sigma}^{-T}+\frac{N}{2}\mathbf{\Sigma^{-1}S\Sigma^{-1}}=^{set}\mathbf{0}$$

and this yields

$$\mathbf{\Sigma}_{ml}=\mathbf{S}=\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n}-\mu)(\mathbf{x_n}-\mu)^T$$

References

- https://atmos.washington.edu/~dennis/MatrixCalculus.pdf
- https://www.comp.nus.edu.sg/~cs5240/lecture/matrix-differentiation.pdf
- https://math.stackexchange.com/questions/1151569/how-to-calculate-the-derivative-of-log-det-matrix
- https://math.stackexchange.com/questions/38701/how-to-calculate-the-gradient-of-log-det-matrix-inverse
- http://www.math.ucdenver.edu/~esulliva/Calculus3/Taylor.pdf
- https://www.quora.com/What-is-the-Jacobian-how-does-it-work-and-what-is-an-intuitive-explanation-of-th
- Pattern Recognition and machine learning, Christopher M. Bishop
- Advanced Engineering Mathematics, Erwin Kreyszig