Linear algebra for computational statistics IV

Jong-June Jeon

September, 2022

Department of Statistics, University of Seoul

Things to know

- basic operation of matrix
- spanning space, null space
- projection and geometry
- linear map and matrix

Matrix Calculus

Differentiation w.r.t vector or matrix

Step 4

$f : \mathbb{R}^p \mapsto \mathbb{R}$ 혹은 $f : \mathbb{R}^{p\times p} \mapsto \mathbb{R}$ 함수의 편미분에 대한 공식을 익힌다. 벡터 혹은 행렬에 대한 미분이 계산을 간단하게 해주며, 주어진 식을 간결하게 표현해 줌을 이해한다. 여기서는 Optimization 자주 사용되는 특별한 형태의 f의 미분 공식을 소개한다. 여기서 소개하는 미분공식의 적용 예를 반드시 직접 만들어보고 확인한다.

Determinant

The determinant of A is the volume of parallelogram in $\mathbb{R}^p$ derived of A:

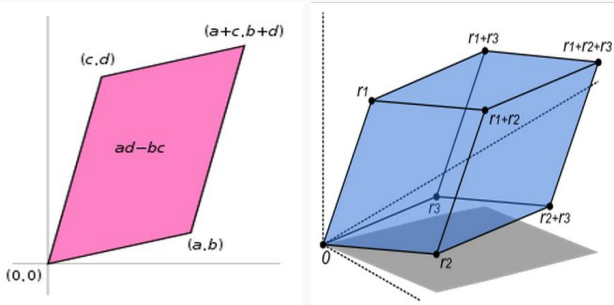

Figure 1: Geometrical meaning of determinant

Determinant

If vectors in A are linearly dependent, the volume of parallelogram is zero, and vice versa.

The following statement is equivalent. Let A be square matrix.

- $det(A) \neq 0$;
- A is full rank;
- $a_1, \cdots, a_p$, column vector of A , are linearly independent.

Properties of determinant

A, B are $p \times p$ matrix.

- $det(A) = det(A^\top)$
- $det(cA) = c^p det(A)$
- $det(AB) = det(A)det(B)$
- If $A = diag(a_1, \cdots, a_p)$, then $det(A) = \prod_{i=1}^{p} a_i$.

Let $f : x \in \mathbb{R}^p \mapsto a^\top x$

$$\frac{\partial f(x)}{\partial x} = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_p} \right)^\top = a$$

Let $f : x \in \mathbb{R}^p \mapsto x^\top Ax$ where $A \in \mathbb{R}^{p \times p}$

$$\frac{\partial f(x)}{\partial x} = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_p} \right)^\top = (A + A^\top)x$$

Let $f : X \in \mathbb{R}^{p\times p} \mapsto \mathsf{Tr}(XA)$ where $A \in \mathbb{R}^{p\times p}$

$$\frac{\partial f(X)}{\partial X} = \begin{pmatrix} \frac{\partial f(x)}{\partial x_{11}}, \cdots, \frac{\partial f(x)}{\partial x_{1p}} \\ \cdots \\ \frac{\partial f(x)}{\partial x_{p1}}, \cdots, \frac{\partial f(x)}{\partial x_{pp}} \end{pmatrix} = A^\top$$

- $f : X \in \mathbb{R}^{p\times p} \mapsto \mathsf{Tr}(X^\top A)$ where $A \in \mathbb{R}^{p\times p}$

$$\frac{\partial f(X)}{\partial X} = A$$

- $f : X \in \mathbb{R}^{p\times p} \mapsto \mathsf{Tr}(AX^{-1}B)$ where $A, B \in \mathbb{R}^{p\times p}$

$$\frac{\partial f(X)}{\partial X} = -(X^{-1}BAX^{-1})^\top$$

- $f : X \in \mathbb{R}^{p\times p} \mapsto \log(det(X))$

$$\frac{\partial f(X)}{\partial X} = X^{-1}$$

(See the matrix cookbook)

Solving linear equations

efficient computation with matrix decomposition

Step 5

closed form의 solution 을 얻을 수 있는 비선형함수는 많지 않다. 그 중 대표적인 것이 Quadaratic function 이다. 대표적으로 $f : x \in \mathbb{R} \mapsto ax^2 + bx + c$, $(a > 0)$는 미분이 0이 되는 값을 구하여 최소값을 찾는다. Quadaratic function의 미분이 0이 되는 해를 찾는 문제는 선형방정식을 문제가 되며, 계산에서는 선형방정식을 효율적을 푸는 것이 중요한 이슈였다. 여기서는 선형방정식을 효율적으로 풀기 위한 행렬의 분해를 배운다. 소개한 행렬의 분해가 어떻게 방정식을 푸는데 유용하게 사용될 수 있는지를 이해한다.

QR decomposition

Let A be $n \times n$ squared matrix. Then,

$$A = \mathbf{QR},$$

where $\mathbf{Q}$ is orthogonal matrix, and $\mathbf{R}$ is upper triangular matrix.

application

Assume that A is invertible and consider a linear system

$$A\mathbf{x} = \mathbf{b}.$$

Then, the system is written by $\mathbf{Rx} = \mathbf{Q}^\top \mathbf{b}$ and the solution $\mathbf{x}$ is easily obtained by iterative computation from x_1 to x_n.

LU decomposition

Let A be $n \times n$ squared matrix. Then,

$$A = \mathbf{L}\mathbf{U},$$

where $\mathbf{L}$ is lower triangular matrix and $\mathbf{U}$ is upper triangular matrix.

remark) Not always exists. (see LUP decomposition)

application

Assume that A is invertible and consider a linear system

$$A\mathbf{x} = \mathbf{b}.$$

Then, the system is written by $\mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{b}$ and the solution $\mathbf{x}$ is easily obtained by forward elimination and backward substitution.

Cholesky decomposition

Let A be $n \times n$ a real-valued symmetric and positive definite matrix. Then,

$$A = \mathbf{L}\mathbf{L}^\top,$$

where $\mathbf{L}$ is a lower triangular matrix.

application

Consider a linear system

$$A\mathbf{x} = \mathbf{b}.$$

Then, the system is written by $\mathbf{L}\mathbf{L}^*\mathbf{x} = \mathbf{b}$, and the solution $\mathbf{x}$ is easily obtained by forward elimination and backward substitution.

What you should know

Step 1

- 행렬과 벡터를 이용하여 여러개의 선형방정식을 간단하게 표현할 수 있다.
- 행렬 위에서 정의된 연산으로 행렬의 덧셈과 곱셈이 있다.
- Transpose, Inverse matrix의 정의와 성질을 이해해야 한다.
- 블록행렬의 계산에 대해 이해하고 있어야 한다.

Step 2

- 행렬이 선형변환과 같음을 설명할 수 있어야 한다.
- 행렬의 곱이 선형변환의 합성임을 설명할 수 있어야 한다.
- 행렬의 열이 span하는 공간과 행렬의 rank의 개념을 설명할 수 있어야 한다.

Step 3

- 내적의 의미를 정사형과 연관하여 설명할 수 있어야 한다.
- 열공간 위에 정사형의 의미를 설명할 수 있고, 그 정사형을 만들어주는 Projection operator를 알아야 한다.
- Eigendecomposition 을 이용하여 행렬을 분해하고 그 의미를 선형변환과 연결하여 설명할 수 있어야 한다.
- SVD에서 사용되는 Orthogonal matrix를 Eigendecomposition을 통해 어떻게 유도할 수 있는지 설명할 수 있어야 한다.

Step 4

- 다양한 형태의 벡터, 행렬 미분을 적용할 수 있어야 한다.

Step 5

- 선형방정식의 해를 찾을 때 많이 사용하는 행렬의 분해를 기억하고, 그것이 왜 사용되는지를 설명할 수 있어야 한다.

Miscellany*

Example: signal recovery (naive upperbound) Suppose that D_{ii} is sorted by descending order, and let $\lambda_i = D_{ii}$. Let $\lambda_{k-1} \geq \epsilon \geq \lambda_k$ and let $\hat{D}$ be $n \times p$ diagonal matrix with

$$\hat{D}_{ii} = \begin{cases} \lambda_i & \text{if } 1 \leq i \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Define $\hat{A} = U\hat{D}V^\top$ then $\hat{A}$ is an approximation of A in the sense that $A\mathbf{x} \simeq \hat{A}\mathbf{x}$.

$$
\begin{aligned}
\|A\mathbf{x} - \hat{A}\mathbf{x}\| &= \|U_k \lambda_k \mathbf{v}_k^\top \mathbf{x} + \cdots + U_p D_{pp} \mathbf{v}_p^\top \mathbf{x}\| \\
&\leq \epsilon \sum_{j=k}^{p} \|U_j \mathbf{v}_j^\top \mathbf{x}\| \text{ (triangular inequality)} \\
&= \epsilon \sum_{j=k}^{p} |(\mathbf{v}_j^\top \mathbf{x})| \|U_j\| \text{ (norm properties)} \\
&\leq \epsilon \sum_{j=k}^{p} \|\mathbf{x}\| \|\mathbf{v}_j\| \|U_j\| \text{ (cauchy inequality)} \\
&= \epsilon(p - k + 1)\|\mathbf{x}\| \text{ (orthogonal matrix)}
\end{aligned}
$$

Example: spectral norm Consider the problem to obtain

$$y = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

Let λ_j be eigenvalue of $A^\top A$ and e_j be the corresponding eigenvector to λ_j. Let $\mathbf{x} = \sum_{j=1}^p a_j e_j$ with $\sum_{j=1}^p a_j^2 = 1$ (let $\|\mathbf{x}\| = 1$), then

$$\begin{aligned}
\|A\mathbf{x}\|^2 &= \mathbf{x}^\top A^\top A\mathbf{x} = \mathbf{x}^\top (\sum_{j=1}^p \lambda_j e_j e_j^\top)\mathbf{x} \\
&= \sum_{j=1}^p \lambda_j (\underbrace{e_j^\top \mathbf{x}}_{=a_j})^\top (e_j^\top \mathbf{x}) = \sum_{j=1}^p \lambda_j a_j^2 \leq (\max_j \lambda_j) \underbrace{\sum_{j=1}^p a_j^2}_{=1}
\end{aligned}$$

Without loss of generality, let $\max_j \lambda_j = \lambda_1$ then the equality holds for $\mathbf{x} = e_1$, which implies $y = \sqrt{\max_j \lambda_j}$ (called of spectral norm of A).

Example: signal recovery (conti.)

$$
\begin{aligned}
& (A-\hat{A})^\top (A-\hat{A}) = (\mathbf{U}(\mathbf{D}-\hat{\mathbf{D}})\mathbf{V}^\top)^\top (\mathbf{U}(\mathbf{D}-\hat{\mathbf{D}})\mathbf{V}^\top) \\
= \quad & \mathbf{V}(\mathbf{D}-\hat{\mathbf{D}})^\top (\mathbf{D}-\hat{\mathbf{D}})^\top \mathbf{V}^\top,
\end{aligned}
$$

which means that the largest eigenvalue of $(A-\hat{A})^\top(A-\hat{A})$ is less than ϵ^2. Hence, the spectral norm of $(A-\hat{A})$ is less than ϵ. and we conclude that $\|(A-A^*)\mathbf{x}\| \leq \epsilon\|\mathbf{x}\|$. Keeping mind of $p \to \infty$ (high dimensional case), compare the result with the previous example (HW).

Example: singular matrix

Let A be symmetric $p \times p$ nonnegative definite matrix. Then the following statements are equivalent:

- A is positive definite.
- the minimum eigenvalue of A is positive.
- A is invertible.
- $det(A) \neq 0$

When $\det(A) = 0$ then we call A is singular.

Let $\mathbf{E}$ be $p \times p$ orthogonal matrix. Then, $\det(\mathbf{E})^2 = 1$ since $\det(\mathbf{E})^2 = \det(\mathbf{E})\det(\mathbf{E}^\top) = \det(\mathbf{E}\mathbf{E}^\top) = \det(I) = 1$. By eigendecomposition $A = \mathbf{E}D\mathbf{E}^\top$, then

$$\det(A) = \det(\mathbf{E}\mathbf{D}\mathbf{E}^\top) = \det(\mathbf{E})\det(D)\det(\mathbf{E}^\top) = \prod_{j=1}^{p} \lambda_j$$

Note that the geometric meaning of determinant is volume. If A is covariance matrix, then $\det(A)$ can be regarded as the volume of A. Moreover, each eigenvalue of λ_j plays a role of the length of edge of p-dimensional rectangle.

Example: Generalized Inverse Matrix

Consider a linear system

$$A\mathbf{x} = \mathbf{y},$$

where A is $n \times m$ matrix and $\mathbf{y}$ is n dimensional column vector. If A is invertible, then $\mathbf{x} = A^{-1}\mathbf{y}$. When, however, A is not invertible (singular or $n \neq m$), how do we figure out the solution of the linear system?

- The solution is a linear map of y through $\mathcal{L} : \mathbb{R}^n \mapsto \mathbb{R}^m$.
- y is the image of A.

Idea of constructing a generalized inverse

If A is invertible ($m = n$), then the solution of the linear system is the image of linear map $\mathcal{L} : \mathbb{R}^n \mapsto \mathbb{R}^m$ of $\mathbf{y}$. So, we expect that there exists $m \times n$ matrix such that $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$. Denote the solution by $\hat{\mathbf{x}}$ and let G be $m \times n$ matrix. Then,

$$\mathbf{A}\hat{\mathbf{x}} = \mathbf{y} \Rightarrow \mathbf{G}\mathbf{A}\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}.$$

Suppose that $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}$, then we also obtain

$$\mathbf{A}(\mathbf{G}\mathbf{A}\hat{\mathbf{x}}) = \mathbf{A}\hat{\mathbf{x}} = \mathbf{y}.$$

Here, we know that $\mathbf{G}\mathbf{A}\hat{\mathbf{x}}$ is the solution of $\mathbf{A}\mathbf{x} = \mathbf{y}$ and $\mathbf{G}\mathbf{A}\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$. Therefore, if we find such a $\mathbf{G}$ then $\mathbf{G}\mathbf{y}$ is a solution of the linear system.

Generalized inverse matrix
Let $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ where $\mathbf{G}$ is $m \times n$ matrix. If $\mathbf{AGA} = \mathbf{A}$, then $\hat{\mathbf{x}}$ is a solution of the linear system.

The $\mathbf{G}$ is called of the generalized inverse matrix of A, and it may not be unique. But the Moore-Penrose pseudo-inverse matrix, a special version of generalized inverse, is unique.

Properties of generalized Inverse

Let G be a generalized inverse of X^TX. The following statements hold.

- $G^\top$ is also a generalized inverse of X^TX;
- $XGX^TX = X$; i.e. GX^T is a generalized inverse of X;
- XGX^T is invariant to G;
- XGX^T is symmetric, whether G is or not.

When G is the penrose inverse of X^TX and X^+ is the penrose inverse of X, then $GX^T = X^+$ (see the reduction of Hermitian case in the penrose inverse matrix).

example

Let $X^TX\boldsymbol{\beta} = X^Tz$ and X^+ is the penrose inverse of X. Then,

$$X\boldsymbol{\beta} = XX^+z.$$

Proof) Let $(X^TX)^+$ be the penrose inverse matrix. Then

$$X(X^TX)^+X^TX\boldsymbol{\beta} = X(X^TX)^+X^Tz.$$

The LHS is $X((X^TX)^+X^T)X\boldsymbol{\beta} = XX^+X\boldsymbol{\beta} = X\boldsymbol{\beta}$ and the RHS is $X((X^TX)^+X^T)z = XX^+z$.