Linear algebra for computational statistics III

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

Decomposition of matrix

Decomposition of linear maps

Step 3

행렬이 대응시키는 변환을 분해하는 과정을 소개한다. 먼저 내적을 도입하여, 벡터공간 위에서 거리와 각도가 자연스럽게 정의되는 과정을 살펴본다. 다음으로 대칭인 반양정치행렬의 분해를 특별한 직교 선형변환의 분해로 이해할 수 있으며, 이를 통해 행렬의 대응을 분해하여 해석한다. 여기서는 내적공간(inner product space)와 정사형(projection), Spectral Decomposition, Singular Value Decomposition을 배운다.

Inner product An inner product space is a vector space V with an inner product:

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

that satisfies the following three properties for all vectors $x, y, z \in V$ and all scalars $a \in \mathbb{R}$.

- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$
- Linearity:

$$\begin{aligned} \langle ax, y \rangle &= a\langle x, y \rangle \\ \langle x + y, z \rangle &= \langle x, z \rangle + \langle y, z \rangle \end{aligned}$$

- Positive-definite: $\langle x, x \rangle > 0, x \in V - \{0\}$

example

Suppose that $a, b \in \mathbb{R}^p$

- Let $\langle a, b \rangle = a^\top b$. Then the $\langle \cdot, \cdot \rangle$ is inner product?
- Let $H \in \mathbb{R}^{p \times p}$ is symmetric and $\langle a, b \rangle = a^\top H b$. Then the $\langle \cdot, \cdot \rangle$ is inner product?
- If H is positive definite, ...

(NOTE) Vector space에는 Addition과 scalar multiplication 연산만 정의되어 있다. Vector space 위에 inner production 연산을 정의해놓으면, Vector space 위에 각도를 정의할 수 있다. 한편 inner production 연산이 주어지면 원소의 길이(norm) 혹은 두 원소간의 거리 (distance)를 정의할 수 있다.

- For $x, y \in V$ define $\langle x, y\rangle = x^\top y$. If $x^\top y = 0$ we write $x \perp y$
- We define the norm of $x \in V$ by $\|x\| = \sqrt{x^\top x}$
- We can define the distance between x and y by $d(x, y) = \|x - y\|$

Hereafter, we use the above definition of the inner product and the norm in our vector space V.

angle and inner product (law of cosine)

- Let a point A, B, C in $\mathbb{R}^2$ and C is the origin and B is a point on x-axis.
- Let the length of $\overline{\mathrm{AB}}$, $\overline{\mathrm{BC}}$ and $\overline{\mathrm{CA}}$ be c, a, and b, respectively.
- Let the angle $\angle C$ be θ.
- The point A is $(b\cos\theta, b\sin\theta)$, and the point $B = (a, 0)$. Thus,

$$
\begin{aligned}
c^2 &= (b\cos\theta - a)^2 + b^2\sin^2\theta \\
&= a^2 + b^2 - 2ab\cos\theta
\end{aligned}
$$

angle and inner product

Because the law of cosine is the fact derived only from geometry, we can apply the law of cosine to a Euclidean space.

Consider a vector u, v, and $u - v$ and denote the norms of the vectors by $\|u\|$, $\|v\|$, and $\|u - v\|$, respectively. Note that $\|u\|$, $\|v\|$, and $\|u - v\|$ correspond to b, a, and c. By the law of cosine

$$\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\|u\|\|v\| \cos(\theta),$$

which reduces to

$$u^\top v = \|u\|\|v\| \cos(\theta).$$

As a result the angle in $\mathbb{R}^p$ are defined by the law of cosine.

angle and inner product Thus,

- $\cos(\theta) = u^\top v / (\|u\|\|v\|)$
- $u^\top v = 0$ is regarded as $u \perp v$

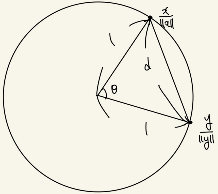
ch) Let u and v be points on a unit sphere and let d be a Euclidean distance between u and v. Then,

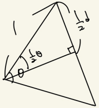
$$u^\top v = \cos(\theta) = 1 - \frac{1}{2} d^2.$$

The equation shows the relationships of inner product, cosine similarity, and distance.

angle and inner product

$$\cos\theta = \left(\frac{x}{\|x\|}\right)^T \left(\frac{y}{\|y\|}\right)$$





$$\sin\frac{\theta}{2} = \frac{1}{2}d$$

$$\Rightarrow \frac{1-\cos\theta}{2} = \frac{1}{4}d^2$$

$$\Rightarrow \cos\theta = 1 - \frac{1}{2}d^2$$

Projection

Suppose that V is an inner product vector space. Let $x, y \in V$ then there exists $\hat{y} \in \mathcal{C}(x)$ such that $(y - \hat{y}) \perp x$. That is y is decomposed into $y = \hat{y} + (y - \hat{y})$ with $\hat{y} \perp (y - \hat{y})$.

Transpose and projection (figure will be corrected!)

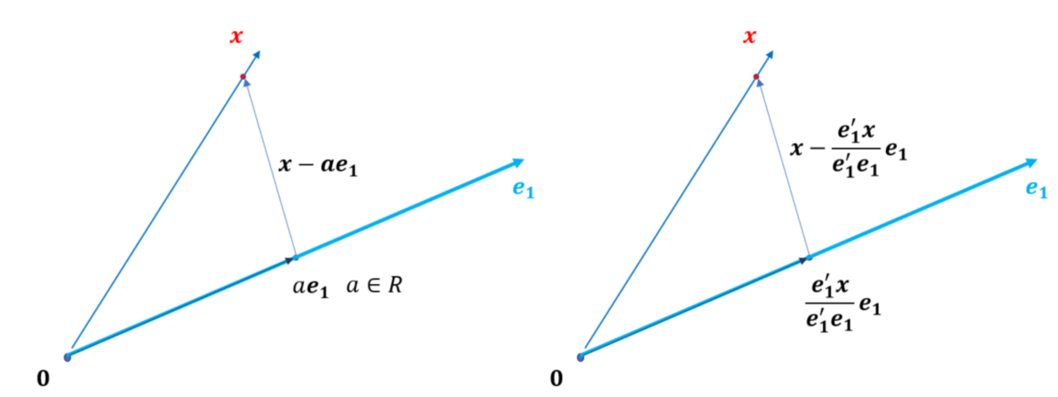


Figure 1: Illustration of projection via transpose operation

$y = ax + (y - ax)$ with $a = y^\top x/\|x\|^2 \in \mathbb{R}$

Projection Let $y, x_1, \cdots, x_k \in V$ and suppose that $x_1, \cdots, x_k$ are linearly independent. Consider $X = [x_1, \cdots, x_k]$ and $\mathcal{C}(X)$. Then, how can we find $\hat{y} \in \mathcal{C}(X)$ such that

$$y = \hat{y} + (y - \hat{y})$$

satisfying $\hat{y} \perp (y - \hat{y})$?

The answer is the Projection map (matrix)!

Orthogonal Projection Let $x, y \in \mathbb{R}^p$ and consider Π, a linear map from $\mathbb{R}^p$ to $\mathbb{R}^p$. The projection map of y onto $\mathcal{C}(x)$ satisfies the following properties:

- $\Pi y \in \mathcal{C}(x)$;
- $(\Pi y)^\top y = y^\top (\Pi y) = 0$;
- $\Pi(\Pi y) = \Pi y$

Orthogonal Projection

Orthogonal projection is a linear map defined by $\Pi : V \mapsto V$ (in fact, we can understand the P as a matrix) such that $\Pi = \Pi^2 = \Pi^\top$.

- Let $\Pi = x(x^\top x)^{-1}x^\top$ then $\Pi = \Pi^2 = \Pi^\top$?
- $\Pi y \in \mathcal{C}(x)$ for $x \in V$?
- $\langle \Pi y, y - \Pi y \rangle = 0$?

You can conclude that the Π is the orthogonal projection operator (onto $\mathcal{C}(x)$).

Orthogonal Projection

- Let $x, y \in \mathbb{R}^n$ and compute the projection of y onto $\mathcal{C}(x)$. Then, it is given by $(x(x^\top x)^{-1}x^\top)y$. Show that $x(x^\top x)^{-1}x^\top$ is a projection operator.
- Let $X \in \mathbb{R}^{n\times m}$ and $y \in \mathbb{R}^n$. Compute the projection of y onto $\mathcal{C}(X)$.
- Write an example and confirm the result numerically.

잠깐!

회귀모형의 OLS를 돌이켜보자. Response vector $Y \in \mathbb{R}^n$과 predictor matrix $X \in \mathbb{R}^{n \times p}$ 일때, OLS를 이용한 Y값의 추정량은

$$\hat{Y} = X(X^\top X)^{-1} X^\top Y \in \mathbb{R}$$

로 주어진다. 여기서 $X(X^\top X)^{-1} X^\top$가 $\mathcal{C}(X)$에 Projection operator 고 $\hat{Y} \in \in \mathcal{C}(X)$ 이며 $(Y - \hat{Y}) \perp \hat{Y}$임을 알 수 있다.

Projection은 벡터 성분을 직교분해할 때 흔히 볼 수 있었던 연산이다.

Basic on undergraduate levels

- Symmetric Matrix
- Orthogonal Matrix

Orthogonal matrix

- Orthogonal matrix E: a square matrix satisfying

$$E = [e_1, \cdots, e_p],$$

  where $e_j^\top e_k = 0$ for $j \neq k$ and $\|e_j\| = 1$ for all j.
- It is easily shown that $E^\top E = I$.
- Because $E(E^\top E) = E$, $EE^\top = I$.

That is, $E^\top E = EE^\top = I$, and $E^\top$ is the inverse of E.

Orthogonal matrix (Isometric transformation)

Let $E \in \mathbb{R}^{p \times p}$ be orthogonal matrix and $x, y \in \mathbb{R}^p$. $d(x, y) = d(Ex, Ey)$?

$$\begin{aligned} d(x,y)^2 &= (x-y)^\top (x-y) = (x-y)^\top E^\top E(x-y) \\ &= \|E(x-y)\|^2 = d(Ex, Ey)^2 \end{aligned}$$

The map $\mathcal{L}_E$ preserves the distance (isometric). Actually, E is understood as the rotation map.

Orthogonal matrix (Rotation)

What is the geometrical meaning of the first column of an orthogonal matrix?

Let E be the orthogonal and $a_1 = (1, 0, \cdots, 0)$.

- Ea_1 is the first column vector.
- In addition, Ea_1 is the image of a linear map E for a_1, which is the first coordinate basis vector.
- That is, the first column vector is the transformed image of the first coordinate basis vector.

To sum up, each column of E denotes an image of each coordinate basis transformed by E. Since the transformed image is orthogonal to each other, the map can be regarded as a geometrical rotation.

Diagonal matrix

Let $D = diag(d_1, \cdots, d_p) \in \mathbb{R}^{p \times p}$ be diagonal matrix and $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$. Then,

$$Dx = (d_1 x_1, \cdots, d_p x_p)^\top$$

The map D is called the scaling map.

앞다행, 뒤다열!

Eigendecomposition

Let $A \in \mathbb{R}^{p \times p}$ be a symmetric matrix. Then there exists an orthogonal matrix E and a diagonal matrix D (with real-valued elements) such that

$$A = EDE^\top$$

- Orthogonality of E: write

  $$E = [e_1, \cdots, e_p]$$

  then $e_j^\top e_k = 0$ for $j \neq k$ and $\|e_j\| = 1$ for all j.
- Projection onto $\mathcal{C}(e_j)$ is given by $e_j(e_j^\top e_j)^{-1} e_j^\top = e_j e_j^\top$

Eigendecomposition Suppose that A be a symmetric matrix. Let λ_j be the jth diagonal element of D, then we can write

$$A = EDE^\top = \sum_{j=1}^{p} \lambda_j e_j e_j^\top$$

We can know that A is the sum of orthogonal projection operators. e_js are eigenvector and λ_j is the associated eigenvalue. $\mathcal{C}(e_j)$ is eigenspace spaned by e_j.

For simplicity let A be 2×2 matrix.

- Let $D_1 = \text{diag}(\lambda_1, 0)$ and $D_2 = \text{diag}(\lambda_2, 0)$, then

$$D_1 E^\top = \lambda_1 \begin{pmatrix} e_1^\top \\ 0 \end{pmatrix} \text{ and } D_2 E^\top = \lambda_2 \begin{pmatrix} 0 \\ e_2^\top \end{pmatrix}$$

- We can easily show that

$$\begin{pmatrix} e_1 & e_2 \end{pmatrix} \begin{pmatrix} e_1^\top \\ e_2^\top \end{pmatrix} = e_1 e_1^\top + e_2^\top e_2$$

  Thus,

$$A = EDE^\top = E(D_1 E^\top + D_2 E^\top) = \lambda_1 e_1 e_1^\top + \lambda_2 e_2^\top e_2$$

Eigendecomposition

This eigendecomposition can be viewed as the decomposition of a linear map:

$$\mathcal{L}_A = \sum_{j=1}^{p} \lambda_j \mathcal{L}_{E_j},$$

where $E_j = e_j e_j^\top$.

Note that

- projection onto $\mathcal{C}(e_j)$ is given by $e_j(e_j^\top e_j)^{-1} e_j^\top = e_j e_j^\top$

Therefore,

$$\mathcal{L}_A(x) = \sum_{j=1}^{p} \lambda_j \mathcal{L}_{E_j}(x),$$

where $\mathcal{L}_{E_j}(x)$ is projection onto the jth eigenspace.

Approximation of Linear map

Let $A^{(k)} = \sum_{j=1}^{k} \lambda_j e_j e_j^\top$ then $A^{(k)}$ approximates A?

$$
\begin{aligned}
EDE^\top \mathbf{x} &= [e_1, \cdots, e_p]\mathsf{diag}(\lambda_1, \cdots, \lambda_p) \begin{pmatrix} e_1^\top \\ \vdots \\ e_p^\top \end{pmatrix} \mathbf{x} \\
&= [e_1, \cdots, e_p]\mathsf{diag}(\lambda_1, \cdots, \lambda_p) \begin{pmatrix} e_1^\top \mathbf{x} \\ \vdots \\ e_p^\top \mathbf{x} \end{pmatrix} \\
&= [e_1, \cdots, e_p] \begin{pmatrix} \lambda_1 e_1^\top \mathbf{x} \\ \vdots \\ \lambda_p e_p^\top \mathbf{x} \end{pmatrix} \\
&= \sum_{j=1}^{p} e_j(\lambda_j e_j^\top \mathbf{x}) = (\sum_{j=1}^{p} \lambda_j e_j e_j^\top)\mathbf{x},
\end{aligned}
$$

Eigendecomposition shows the linear map of a symmetric matrix as the composition of three operations:

$$Ax = EDE^\top x$$

$$\begin{aligned} x \mapsto E^\top x \text{ (rotation)} &\mapsto D(E^\top x) \text{ (scaling)} \\ &\mapsto E(DE^\top x) \text{ (reverse rotation)} \end{aligned}$$

Inverse matrix of positive definite matrix

Let $\mathbf{A}$ be symmetric and nonnegative definite matrix. Then the minimum eigenvalue is positive if and only if $\mathbf{A}$ is positive definite.

pf) Let λ_{min} be the minumum eigenvalue of $\mathbf{A}$. Assume that $\lambda_{min} > 0$. Let $\mathbf{x} = \sum_{j=1} a_j e_j \neq 0$, then

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{j=1}^{p} \lambda_j (e_j^\top \mathbf{x})^2 = \sum_{j=1}^{p} \lambda_j a_j^2 > 0.$$

Assume that $\mathbf{A}$ is pd matrix. WLOG, let λ_p be the minimum eigenvalue of $\mathbf{A}$. Then,

$$e_p^\top \mathbf{A} e_p = \sum_{j=1} \lambda_j (e_j^\top e_p)^2 = \lambda_p > 0.$$

Inverse matrix of positive definite matrix

The inverse matrix of such $\mathbf{A}$ is given by

$$\mathbf{A}^{-1} = \mathbf{E}D^{-1}\mathbf{E}^\top.$$

pf) $\mathbf{E}D^{-1}\mathbf{E}^\top\mathbf{A} = \mathbf{E}D^{-1}\underbrace{\mathbf{E}^\top\mathbf{E}}_{=I}D\mathbf{E}^\top = I$

and $\mathbf{A}\mathbf{E}D^{-1}\mathbf{E}^\top = \mathbf{E}D\underbrace{\mathbf{E}^\top\mathbf{E}}_{=I}D^{-1}\mathbf{E}^\top = I$. By definition of the inverse matrix, we obtain the result.

잠깐!

특별히 pd matrix $A \in \mathbb{R}^{p\times p}$에 대해서

$$\lambda_{\min} = \min_x \frac{x^\top Ax}{\|x\|^2}, \quad \lambda_{\max} = \max_x \frac{x^\top Ax}{\|x\|^2}$$

라 놓으면 $\lambda_{\max} \geq \lambda_{\min} > 0$ 이며, $\lambda_{\max}$, $\lambda_{\min}$ 각각 A의 maximum eigenvalue, minimum eigenvalue에 해당한다. 한편 pd matrix A에서 Eigenmatrix의 열 $e_1, \cdots e_p$는 다음과 같이 구할 수 있다.

- $e_1 = \text{argmax}_{x\in\mathbb{R}^p} \frac{x^\top Ax}{\|x\|^2}$
- $e_2 = \text{argmax}_{x\in\mathbb{R}^p : x\perp e_1} \frac{x^\top Ax}{\|x\|^2}$
- $e_3 = \text{argmax}_{x\in\mathbb{R}^p : x\perp \mathcal{C}([e_1,e_2])} \frac{x^\top Ax}{\|x\|^2}$
- $\cdots$

Singular Value decomposition

Let A be $n \times p$ a real valued matrix. Then,

$$A = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix, V is an $p \times p$ orthogonal matrix and D is a $n \times p$ rectangular diagonal matrix where $(D)_{ij} = 0$ for $i \neq j$ and $(D)_{ii} \geq 0$ for $1 \leq i \leq \min(n, p)$.

- $AA^\top = U\bar{D}U^\top$ ($\bar{D} = DD^\top$ is diagonal matrix)
- $A^\top A = V\tilde{D}V^\top$ ($\tilde{D} = D^\top D$ is diagonal matrix)

Thus, U and V are eigenmatrix of $AA^\top$ and $A^\top A$, respectively.

The singular value decomposition gives an insight for understanding of linear map. Let A be $n \times p$ $(n > p)$ matrix which is a linear map from $\mathbb{R}^p$ and $\mathbb{R}^n$.

$$
\begin{aligned}
A\mathbf{x} &= UDV^\top \mathbf{x} = U \begin{pmatrix} d_1 \mathbf{v}_1^\top \mathbf{x} \\ \vdots \\ d_p \mathbf{v}_p^\top \mathbf{x} \\ \mathbf{0_{n-p}} \end{pmatrix} \\
&= \sum_{j=1}^{p} (d_j \mathbf{v}_j^\top \mathbf{x}) U_j
\end{aligned}
$$

A linear map A is interpreted as the compositions of

- projection map onto $\mathcal{C}(V)$ and
- scaling map and
- rotation map (scale invariant map).

Singular Value decomposition: Data representation

$$X \quad = \quad UDV^\top \text{ (Data matrix)}$$

Let $x_i^\top$ and $u_i^\top$ be the ith row vectors of X and U. Since $x_i^\top = u_i^\top DV^\top$,

$$x_i = V(D^\top u_i)$$

Let $d_i = (D)_{ii}$ for $1 \le i \le p$ and $D^\top u_i = (d_1 u_{i1}, \cdots, d_p u_{ip})^\top$. Then

$$x_i = V_1 \beta_1 + \cdots + V_p \beta_p,$$

where $\beta_j = d_j u_{ij}$. Note that $(\beta_1, \cdots, \beta_p)^\top$ be the representation of x_i with respect to $V_1, \cdots, V_p$. Thus, we know that $D^\top u_i$ is the representation of x_i with respect to V.

Singular Value decomposition: Data representation

- Suppose that X is normalized. Then $X^\top X/n$ is the sample covariance matrix.
- PCA employs the eigenvector of the sample covariance, $X^\top X/n$:

$$\Sigma = E\Lambda E^\top = X^\top X/n = V((D^\top D)/n)V^\top$$

That is, V of SVD is equal to E of PCA.