Review: matrices & decompositions

Outer and inner products

Inner product of $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^n$ and $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$ is

$$a^T b = \sum_{i=1}^{n} a_i b_i = \langle a, b \rangle$$

Recall, a and b are **orthogonal** if $a^T b = 0$

Outer product of $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^n$ and $b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$ is

$$a \cdot b^T = \begin{pmatrix} a_1 b_1 & \cdots & a_1 b_m \\ \vdots & & \vdots \\ a_n b_1 & \cdots & a_n b_m \end{pmatrix} \in \mathbb{R}^{n \times m}$$

$\frac{a^T}{\ } \Big| b = \bullet$ scalar

$a \Big| \frac{\ }{b^T} =$ matrix

Inner product vs Outer product

Given $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times m}$ there are two ways to view the product AB.

"inner product view"

$$A = \begin{pmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_n^T & - \end{pmatrix} \qquad B = \begin{pmatrix} | & & | \\ b_1 & \cdots & b_m \\ | & & | \end{pmatrix}$$

$$AB = \begin{pmatrix} a_1^T b_1 & \cdots & a_1^T b_m \\ \vdots & & \vdots \\ a_n^T b_1 & \cdots & a_n^T b_m \end{pmatrix}$$

"outer product view"

$$A = \begin{pmatrix} | & & | \\ a_1 & \cdots & a_p \\ | & & | \end{pmatrix} \qquad B = \begin{pmatrix} - & b_1^T & - \\ & \vdots & \\ - & b_p^T & - \end{pmatrix}$$

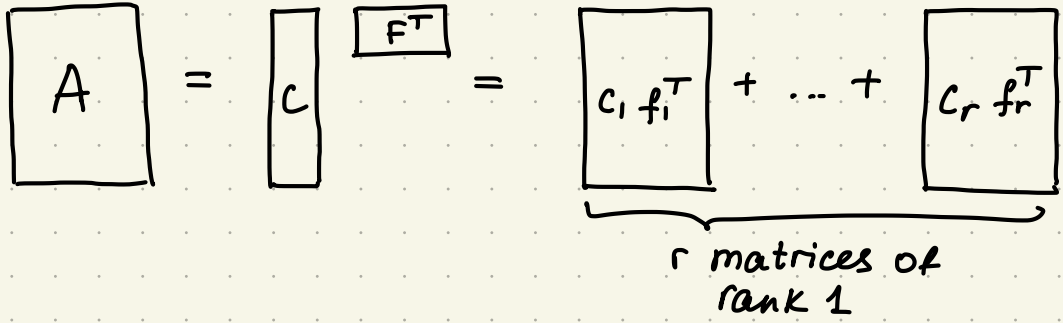
$$AB = \sum_{j=1}^{p} a_j \cdot b_j^T = \underbrace{\square_{n \times m} + \cdots + \square_{n \times m}}_{p \text{ matrices}}$$

Rank factorization

Given $A \in \mathbb{R}^{n \times p}$ with $\text{rank}(A) = r$, one can find

- full column rank $C \in \mathbb{R}^{n \times r}$
- full column rank $F \in \mathbb{R}^{p \times r}$

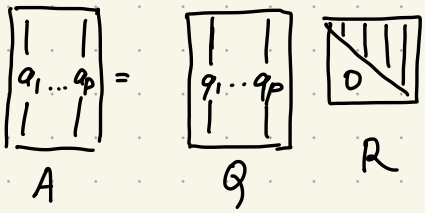
such that $A = C F^T = \sum_{j=1}^{r} c_j f_j^T$



QR decomposition

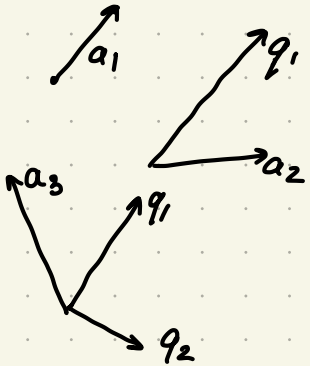
Given $A \in \mathbb{R}^{n \times p}$ it can be decomposed as $A = QR$

where
- Q is column-orthogonal, i.e. $Q^T Q = I$
- R is upper-triangular

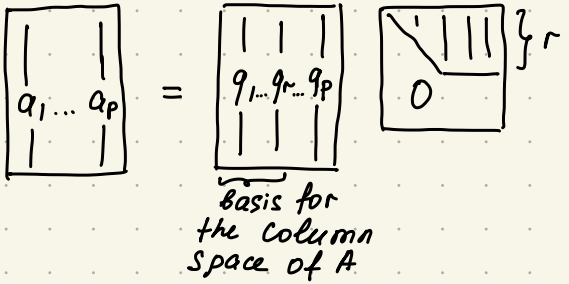


To find Q and R Gram-Schmiat proceedure is used.

- $e_1 = a_1 \longrightarrow q_1 = \frac{e_1}{\|e_1\|}$
- $e_2 = a_2 - \langle a_2, q_1 \rangle q_1 \longrightarrow q_2 = \frac{e_2}{\|e_2\|}$
- $e_k = a_k - \sum_{i=1}^{k-1} \langle a_k, q_i \rangle q_i \longrightarrow q_k = \frac{e_k}{\|e_k\|}$



What if $\text{rank}(A) = r < p$? This will be reflected in R.



Example: linear regression

$$\hat{\beta} = (X^T X)^{-1} X^T y = (R^T R)^{-1} R^T Q^T y = R^{-1} Q^T y$$

$R\hat{\beta} = Q^T y$ can be solved via back solving (fast!)

Eigen (spectral) decomposition

Consider $A \in \mathbb{R}^{n \times n}$, assume A is symmetric.

Then $A = U \Lambda U^T$ where

- $U \in \mathbb{R}^{n \times n}$ is orthogonal, i.e. $U^T U = U U^T = I$
- $\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ is diagonal

$U = \begin{pmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{pmatrix}$, u_i are eigenvectors

λ_i are corresponding eigenvalues

Sometimes, it is useful to view ED as

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

Properties:

- $A u_j = \lambda_j u_j$, i.e. mapping A doesn't rotate u_j.

  $A = \sum_{i=1}^{n} \lambda_i u_i u_i^T \Rightarrow A u_j = \sum_{i=1}^{n} \lambda_i u_i u_i^T u_j = \lambda_j u_j$

- Eigenvectors are determined up to sign

  $u_i \longrightarrow -u_i$

- The rank of A is the number of non-zero λ_i.
- If $V = \sum_{i=1}^{n} \alpha_i u_i$ then $AV = \sum_{i=1}^{n} \lambda_i \alpha_i u_i$

  $AV = \sum_{i=1}^{n} \alpha_i A u_i = \sum_{i=1}^{n} \alpha_i \lambda_i u_i$

  (λ_i: scaling factors)

Definite Matrices

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive-definite (PD)
if $v^T A v > 0$ for all $v \in \mathbb{R}^n$ $v \neq 0$

It is positive semi-definite (PSD)
if $v^T A v \geq 0$ for all $v \in \mathbb{R}^n$

PSD: $A \succcurlyeq 0$ PD: $A \succ 0$

Properties:

- $A \succcurlyeq 0 \iff \lambda_i \geq 0$
- $A \succ 0 \iff \lambda_i > 0$

| $A \succ 0 \Rightarrow v^T A v > 0 \quad \forall v \neq 0$, let $v = u_i$
| $u_i^T A u_i = \lambda_i \|u_i\|^2 = \lambda_i > 0$
| $\lambda_i > 0 \Rightarrow v^T A v = (U^T v)^T \Lambda (\underbrace{U^T v}_{y}) = \sum_{i=1}^{n} \lambda_i (y_i)^2$

- ED can be used to compute powers of A as $A^k = U \Lambda^k U^T$.

| $A^k = U \underbrace{\Lambda \cancel{U^T} \cdot \cancel{U} \Lambda \cancel{U^T} \cancel{U} \Lambda}_{\Lambda^k} U^T = U \Lambda^k U^T$

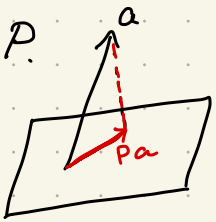
- for $A \succ 0$ it holds $A^{-1} = U \Lambda^{-1} U^T$
  for $A \succcurlyeq 0$ it is true $A^{1/2} = U \Lambda^{1/2} U^T$

| $A^{1/2} \cdot A^{1/2} = U \Lambda^{1/2} \cancel{U^T} \cancel{U} \Lambda^{1/2} U^T = U \Lambda U^T = A$

Projection

Matrix $P \in \mathbb{R}^{n \times n}$ is a **projection matrix** if $P^2 = P$.

Orthogonal projection is when $P = P^T = P^2$.



Properties: if P is orthogonal projection then

- Eigenvalues are 0 or 1

$P = U \Lambda U^T, \quad P^2 = U \Lambda^2 U^T = P = U \Lambda U^T$

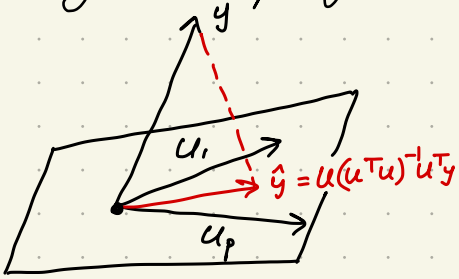
- It is PSD.
- Diagonal elements are bounded $0 \leq p_{ii} \leq 1$

$P = P^2 = P^T P \Rightarrow p_{ii} = \sum_{k=1}^{n} p_{ki}^2 \geq p_{ii}^2 \geq 0$

Examples

- Given $U \in \mathbb{R}^{n \times p}$, $P = U(U^TU)^{-1}U^T$ is orthogonal projection onto the columnspace of U

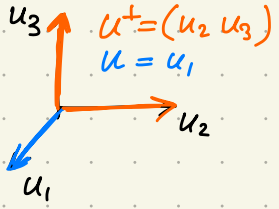
  $y \in \mathbb{R}^n \quad P_U(y) = U\beta$ and minimize $\|y - U\beta\|^2$ — regression, $\beta \in \mathbb{R}^p$



- If U is column orthogonal $P = UU^T$.

- $P = I - UU^T$ is projection onto the column space of $U_\perp$

  $\tilde{U} = (U, U_\perp) \Rightarrow \tilde{U}\tilde{U}^T = UU^T + U_\perp U_\perp^T = I$



$U_\perp$ is orthogonal complement of the basis U if

- $U_\perp \in \mathbb{R}^{n \times (n-p)}$ is column orthogonal, i.e. $U_\perp^T U_\perp = I$
- $\tilde{U} = (U \; U_\perp) \in \mathbb{R}^{n \times n}$ is orthogonal, i.e. $\tilde{U}\tilde{U}^T = \tilde{U}^T\tilde{U} = I$

This also implies that

- $U_\perp$ is column orthogonal and $U^T U_\perp = 0_{p \times (n-p)}$

Column centering

Given sample matrix $X \in \mathbb{R}^{n \times p}$, $X = \begin{pmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{pmatrix}$

- the **sample mean** is $\bar{x} = \frac{X^T 1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- the column centered version is $\tilde{X} = C \cdot X$,

where $C = I - \frac{1 \cdot 1^T}{n} \in \mathbb{R}^{n \times n}$ is **centering operator**

$$CX = X - \frac{1}{n} 1 1^T X = X - 1 \cdot \bar{x}^T = X - \begin{pmatrix} \bar{x}^T \\ \vdots \\ \bar{x}^T \end{pmatrix}$$

C is orthogonal projection operator
(onto what space?)

$u = \frac{1}{\sqrt{n}} \cdot 1_n \in \mathbb{R}^n$ then $u^T u = 1$, C is projection on $u^\perp$

Sample covariance matrix

Given $X = (f_1' \dots f_p') \in \mathbb{R}^{n \times p}$ **sample covariance matrix** is $S \in \mathbb{R}^{p \times p}$ with $S_{ij} = \operatorname{cov}(f_i, f_j)$.

If $X = \begin{pmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{pmatrix}$ and $C = I - \frac{1 1^T}{n}$, then

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T = \frac{X^T C X}{n-1}$$

$$X^T C X = X^T C^T C X = \tilde{X}^T \tilde{X}$$

If X is centered then $S = \frac{X^T X}{n-1}$

Properties:

- Sample covariance matrix is PSD

$$v^T S v = \frac{1}{n-1} v^T X^T X v = \frac{\|Xv\|^2}{n-1}$$

- In fact, $v^T S v = \text{var}(Xv)$
- In regression $\hat{\beta} = S_x^{-1} S_{xy}$ where $S_x \in \mathbb{R}^{p \times p}$ is sample covariance for x and $S_{xy} \in \mathbb{R}^{p \times 1}$ is sample covariance between x and y

$$\hat{\beta} = \left(\frac{X^T X}{n-1}\right)^{-1} \frac{X^T y}{n-1}$$

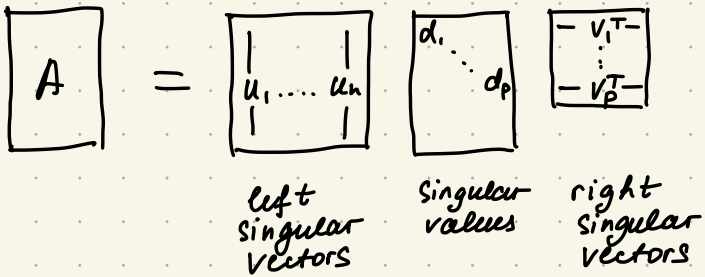
Singular Value Decomposition (SVD)

Given $A \in \mathbb{R}^{n \times p}$, SVD expresses A as

$A = UDV^T$ where

- $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal
- $D \in \mathbb{R}^{n \times p}$ diagonal with non-negative entries sorted in decreasing order.

$n > p$

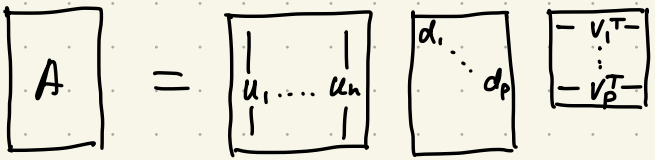


$U^TU = UU^T = I$

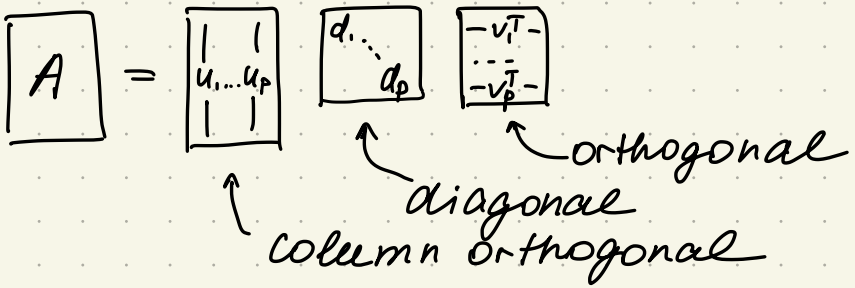
$V^TV = VV^T = I$

$d_1 \geq \dots \geq d_p \geq 0$

$n > p$



Compact form for SVD:



Note that $A = \sum_{i=1}^{\min(n,p)} d_i \cdot u_i \cdot v_i^T$

$u_i \cdot v_i^T$: rank-1 matrices

d_i: ranked by "importance"

Properties:

- $A v_i = d_i u_i$ and $A^T u_i = d_i v_i$ for $i \le \min(n, p)$

  $A v_j = \sum_{i=1}^{p} d_i u_i v_i^T v_j = d_j u_j \|v_j\|^2$

- Every A has SVD.
- D is unique, U and V are not.

  For example, if we have repeated singular values there are many ways to compute U and V.

  $D = \begin{pmatrix} d & & \\ & \ddots & \\ & & d \end{pmatrix} = dI$ then $UDV^T = \underbrace{UQ}\, D\, \underbrace{Q^T V^T} = \tilde{U} D \tilde{V}^T$

  $D = \begin{pmatrix} d & & & & \\ & \ddots & & & \\ & & d & & \\ & & & d_{k+1} & \\ & & & & \ddots \\ & & & & & d_p \end{pmatrix}$ $\quad U = \begin{pmatrix} | & & | & \\ u_1 & \cdots & u_k & \cdots \\ | & & | & \end{pmatrix}$ $\quad V = \begin{pmatrix} | & & | & \\ v_1 & \cdots & v_k & \cdots \\ | & & | & \end{pmatrix}$

  Q $\qquad$ Q

- If $\text{rank}(A) = r$, then $d_{r+1} = \ldots = d_p = 0$

Link between SVD and ED

- If A is PSD, the decompositions are the same.

What if A is symmetric but not PSD?

$A = U \Lambda U^T$ if $\lambda_i \geq 0$ then $U = U$, $D = \Lambda$, $V = U$
If $\lambda_i < 0$ then $u_i = u_i$, $d_i = -\lambda_i$, $v_i = -u_i$

Given $A \in \mathbb{R}^{n \times p}$ with $n > p$

- v_i are eigenvectors for A^TA, d_i^2 are eigenvalues

$A = U D V^T \quad A^TA = V D \cancel{U^T} \cancel{U} D V^T = V D^2 V^T$

(D: $p \times p$)

- u_i are eigenvectors for AA^T, d_i^2 are eigenvalues (plus some extra zeroes)

$AA^T = U D^2 U^T$

(U: $n \times p$, D^2: $p \times p$)

Example $X = UDV^T$ then ridge regression solution

$$\hat{y}_\lambda = X\hat{\beta}_\lambda = X(X^TX + \lambda I)^{-1}X^Ty = \sum_{i=1}^{\min(n,p)} \underbrace{\frac{d_i^2}{d_i^2+\lambda}}_{\text{scaled down factor}} \underbrace{\langle u_i, y\rangle}_{\text{length of projection of } y \text{ onto } u_i} u_i$$

$$\hat{y}_\lambda = UDV^T(VD^2V^T + \lambda I)^{-1}VDU^Ty =$$

$$= UDV^TV(D^2+\lambda I)^{-1}V^TVDU^Ty =$$

$$= U\frac{D^2}{D^2+\lambda I}U^Ty = \sum_{i=1}^{p}\frac{d_i^2}{d_i^2+\lambda}u_iu_i^Ty$$

Matrix and Vector norms

Given vector $V \in \mathbb{R}^n$

- ℓ_1 norm $\|V\|_1 = \sum_{i=1}^{n} |v_i|$
- ℓ_2 norm $\|V\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$ (often $\|V\|$)

Given matrix $A \in \mathbb{R}^{n \times p}$

- Frobenius norm $\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij}^2 = tr(A^T A) = \sum_{i=1}^{\min(n,p)} d_i^2(A)$

Recall that for $A \in \mathbb{R}^{n \times n}$ trace is $tr(A) = \sum_{i=1}^{n} a_{ii}$

Properties

- $tr(A) = tr(A^T)$
- $tr(ABC) = tr(CAB) = tr(BCA)$

- Nuclear norm $\|A\|_* = \sum_{i=1}^{\min(n,p)} d_i(A)$

Gradients

If $f: \mathbb{R}^n \to \mathbb{R}$, so it takes $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ and maps it to a number, then

$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^n$$

If $f: \mathbb{R}^{n \times p} \to \mathbb{R}$, so it takes $A = \begin{pmatrix} A_{11} \dots A_{1p} \\ A_{n1} \dots A_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$ and maps it to a number, then

$$\nabla_A f(A) = \begin{pmatrix} \frac{\partial f(A)}{\partial A_{11}} & \cdots & \frac{\partial f(A)}{\partial A_{1p}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{n1}} & \cdots & \frac{\partial f(A)}{\partial A_{np}} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

Examples:

- $S \in \mathbb{R}^{n \times n}$, $S = S^T$, $x \in \mathbb{R}^n$

  if $f(x) = x^T S x$ then $\nabla_x f(x) = 2Sx$

  $x^T S x = \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij} x_i x_j$ $\quad \frac{\partial f(x)}{\partial x_i} = 2 \sum_{j=1}^{n} S_{ij} x_j = 2 S_i^T x$

  $\nabla_x f(x) = \begin{pmatrix} 2 S_1^T x \\ \vdots \\ 2 S_n^T x \end{pmatrix} = 2Sx$

- $A, B \in \mathbb{R}^{n \times p}$

  if $f(A) = \|A\|_F^2$ then $\nabla_A f(A) = 2A$

  $\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij}^2$ $\quad \frac{\partial f(A)}{\partial A_{ij}} = 2 A_{ij}$ $\quad \nabla_A f(A) = 2 \begin{pmatrix} A_{11} & \dots & A_{1p} \\ \vdots & & \vdots \\ A_{n1} & \dots & A_{np} \end{pmatrix} = 2A$

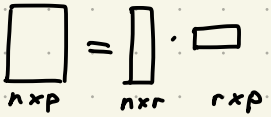
  if $f(A) = tr(AB^T)$ then $\nabla_A f(A) = B$

  $tr(AB^T) = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij} B_{ij}$ $\quad \frac{\partial f(A)}{\partial A_{ij}} = B_{ij}$

Recap from Lecture 1

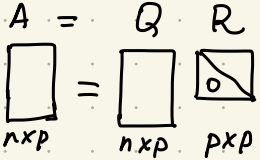
Rank factorization: $\text{rank}(A) = r \Rightarrow C$ and F are full col. rank

$$A = C \cdot F^T$$



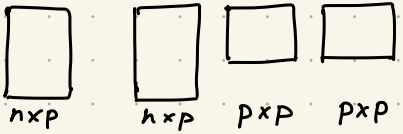
QR decomposition: $Q^TQ = I$, R is upper triangular

$$A = Q\,R$$

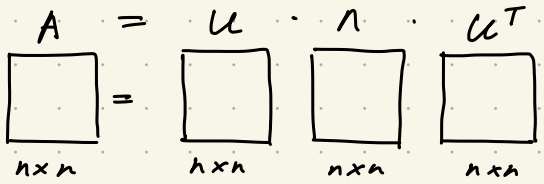


SVD: $U^TU = I$, $D = \text{diag}(d_1, \dots d_p)$, $d_1 \geq \dots \geq d_p \geq 0$, $V^TV = VV^T = I$

$$A = U \cdot D \cdot V^T$$



ED: $A^T = A \Rightarrow U^T U = UU^T = I$, $\Lambda = \text{diag}(d_1, \dots d_n)$



Property: $A = UDV^T \Rightarrow AA^T = \underbrace{UD^2U^T}_{\text{e.vec}\ \ \text{e.val}}$ and $A^TA = \underbrace{VD^2V^T}_{\text{e.vec}\ \ \text{e.val}}$

Orthogonal projection: $P^2 = P = P^T$

- onto $U = (u_1, \dots u_p)$ $\Rightarrow P = U(U^TU)^{-1}U^T$
- onto U if $U^TU = I$ $\Rightarrow P = UU^T$
- onto $U_\perp$ $\Rightarrow P = I - UU^T$

Example: $C = I - \frac{11^T}{n}$ centering operator

PSD: Symmetric $A \succcurlyeq 0$ if $V^TAV \geq 0$ for any V.

equivalently, $\lambda_i(A) \geq 0$

Examples: orth. projection, sample covariance