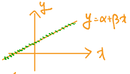
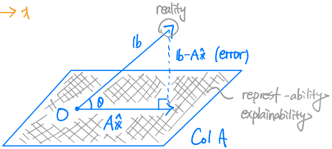
Quadratic Form and Covariance Matrix

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

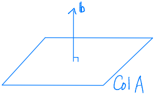
I. Notice & Review

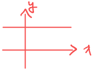
Understanding θ (L6.p11)





- θ can differ, what does it mean?
  - $\theta = 0°$ $\quad b = \hat{b}$
    - Lin. Reg. reflects $\boxed{\text{reality}}$ (perfectly)
    - $\boxed{\text{all points}}$ are on the linear regression line exactly.
  - Small θ
    - $A\hat{x}$ and b are closer, error is small
    - Lin. Reg. reflects reality (well) least-squared
  - Large θ
    - Ax and b are further,
    - Lin. Reg. reflects reality (poorly)
  - $\theta = 90°$ meaningless
    - Lin. Reg. reflects reality (nothing)
    - $y = \alpha + \beta x$, $\beta = 0$.





- $Cos\ \theta = R^2$
  - [R^2] measures explanatory power in percentage term
  - [R^2] measures the percentage of variations explained by linear regression

- One can apply cosine law to find R^2 as well by

$$\|b - A\hat{x}\|^2 = \|A\hat{x}\|^2 + \|b\|^2 - 2\|A\hat{x}\| \cdot \|b\| \cdot \cos\theta$$

II. Quadratic forms and definite matrix

Motivation

$A = PDP^{-1}$ (diagonalization)
$A = PDP^{T}$ (orthogonal diagonalization)

In orthogonal diagonalization at `L7.p4-p5`, we had

$$A = P \underbrace{\begin{bmatrix} \sqrt{7} & & \\ & \sqrt{7} & \\ & & \sqrt{-2} \end{bmatrix}}_{\sqrt{D}} \underbrace{\begin{bmatrix} \sqrt{7} & & \\ & \sqrt{7} & \\ & & \sqrt{-2} \end{bmatrix}}_{\sqrt{D}^{T}} P^t$$

Then, it followed

$$\begin{aligned} A &= P\sqrt{D} \cdot \sqrt{D}P^t \\ &= (P\sqrt{D}) \cdot (P\sqrt{D})^t \end{aligned}$$

- This makes *less sense* (depending on the way you look at) since complex numbers are involved.
- If all eigenvalues (here, 7, 7, -2) were positive real numbers, then it will make more sense!

Definite matrix

concerns with signs of eigenvalues

- **Definition.** A symmetric matrix   is called
  - **positive definite (pd)** if all eigenvalues are positive
  - **positive semi-definite (psd)** if all eigenvalues are non-negative
  - **negative semi-definite (nsd)** if all eigenvalues are non-positive
  - **negative definite (nd)** if all eigenvalues are negative
  - **indefinite** if signs of eigenvalues are mixed

quadratic polynomial of x_1, x_2, x_3
이차 다항식

$$x^T A x = [x_1\ x_2\ x_3] \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

- What makes us to call 'definitely positive'?
  - Since every eigenvalue is positive
  - If A is pd, then $[x_1\ \ x_2\ \ x_3] \cdot A \cdot [x_1\ \ x_2\ \ x_3]^t$ is always positive no matter what $x_1\ \ x_2\ \ x_3$ values are.
  - **This is where second degree polynomial and matrix algebra meet!**
  - The following polynomial is always positive for nonzero x since all eigenvalues are positive (Check the eigenvalues yourself).

pd

$$[x_1\ \ x_2] \underbrace{\begin{bmatrix} 1 & 9 \\ 9 & 100 \end{bmatrix}}_{\lambda_1 \& \lambda_2 > 0} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1x_1^2 + 18x_1x_2 + 100x_2^2 > 0$$

Why is the polynomial positive?

- pd matrix is symmetric, thus orthogonally diagonalizable.
- pd matrix has eigenvalues that are all positive.

y^T $\quad = A = PDP^T \quad$ y

orthonormal eigenvectors

eigen values

$$x^TAx$$

$$= [x_1\ x_2\ x_3] \begin{bmatrix} | & | & | \\ \boxed{u_1 \quad u_2 \quad u_3} \\ | & | & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix} \begin{bmatrix} - & u_1 & - \\ - & u_2 & - \\ - & u_3 & - \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Letting y $= U^t$x gives

$$= [y_1\ y_2\ y_3] \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix} \boxed{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \lambda_3 y_3^2$$

- Since all λ_i are positive and y_i are real numbers, the above polynomial is positive.
- Of course, this applies to all classes of the other definite matrices (pd, psd, nd, nsd) as well.

Applications in optimization

- Linear Programming
  - Objective function & constraints $\rightarrow$ both linear
- Non-Linear Programming
  - Semi-definite programming
    - Objective function & constraints $\rightarrow$ semi-definite polynomial or linear
    - Some are introduced in our textbook
  - Other non-linear programming
    - Problems in this class are incredibly hard to solve

III. Covariance matrix & Principal component analysis

Covariance matrix

- Covariance matrix is a representative example of psd.
  - Covariance matrix is symmetric, thus orthogonally diagonalizable (`Theorem 2 in Section 7.1`)
  - Covariance matrix is psd non-negative, since a variance of linear combination of random variable is always nonnegative. (Related fields include multivariate statistics and portfolio theory)

$$Cov = \Sigma = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Cov(X_3, X_3) \end{bmatrix}$$

Orthogonal diagonalization on psd

Since covariance matrix is symmetric, thus being orthogonally diagonalizable, let's do one with a sample covariance matrix S. Assume that eigenvalues are known as: $\lambda_1 = 9, \lambda_2 = 6, \lambda_3 = 3,$

$$S = \begin{bmatrix} 5 & 2 & 0 \\ 2 & 6 & 2 \\ 0 & 2 & 7 \end{bmatrix}$$

i) $\lambda_1 = 9$, $S - \lambda_1 I = \begin{bmatrix} -4 & 2 & 0 \\ 2 & -3 & 2 \\ 0 & 2 & -2 \end{bmatrix}$

$\mathbb{V}_1 = [1 \;\; 2 \;\; 2]^T$

$\mathbb{U}_1 = \frac{1}{3}[1 \;\; 2 \;\; 2]^T$

ii) $\lambda_2 = 6$, $S - \lambda_2 I = \begin{bmatrix} -1 & 2 & 0 \\ 2 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix}$

$\mathbb{V}_2 = [2 \;\; 1 \;\; -2]^T$

$\mathbb{U}_2 = \frac{1}{3}[2 \;\; 1 \;\; -2]^T$

iii) $\lambda_3 = 3$ $S - \lambda_3 I = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 3 & 2 \\ 0 & 2 & 4 \end{bmatrix}$

$\mathbb{V}_3 = [2 \;\; -2 \;\; 1]^T$

$\mathbb{U}_3 = \frac{1}{3}[2 \;\; -2 \;\; 1]^T$

$$S = PDP^T, \quad P = \begin{bmatrix} | & | & | \\ \mathbb{u}_1 & \mathbb{u}_2 & \mathbb{u}_3 \\ | & | & | \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{bmatrix}$$

$$S = \begin{bmatrix} | & | & | \\ \mathbb{u}_1 & \mathbb{u}_2 & \mathbb{u}_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} 9 & & \\ & 6 & \\ & & 3 \end{bmatrix} \begin{bmatrix} - & \mathbb{u}_1^T & - \\ - & \mathbb{u}_2^T & - \\ - & \mathbb{u}_3^T & - \end{bmatrix}$$

Principal component analysis (PCA)

From the previous example, we have

$= 9 \cdot u_1 \cdot u_1^T + 6 \cdot u_2 \cdot u_2^T + 3 \cdot u_3 \cdot u_3^T = \sum \lambda_i \cdot u_i \cdot u_i^T$

$$Cov = S = PDP^t = \begin{bmatrix} | & | & | \\ u_1 & u_2 & u_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} 9 & & \\ & 6 & \\ & & 3 \end{bmatrix} \begin{bmatrix} - & u_1^T & - \\ - & u_2^T & - \\ - & u_3^T & - \end{bmatrix},$$

where

$$u_1 = \frac{1}{3}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \ u_2 = \frac{1}{3}\begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}, \ u_3 = \frac{1}{3}\begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$$

- When a covariance matrix went throught orthogonal diagonalization, we call u_1, u_2, u_3 as **principal components(PC)** of original data. orthonormal eigenvector
- The first PC u_1 explains $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{9}{9+6+3} = 50\%$ of overall variation
- The second PC u_2 explains $\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = \frac{6}{9+6+3} = 33\%$ of overall variation
- The third PC explains 17% of overall variation

- Though the original data had three variables, the first two PCs (u_1 and u_2) explains 83% of overall variation.
- The remaining third PC explains only 17% of overall variation
- If ignoring third PC, dimension would be reduced into two, but information loss is only 17%
- PCA is one of dimension reduction techniques and popular these days due to big data with a lot of variables.
- In statistical learning field, PCA is one of unsupervised learning methods.
- (google PCA on `mnist` if you like)
- **Conducting PCA is nothing but doing orthogonal diagonalization on a covariance matrix**

IV. pd matrix & Cholesky decomposition

Cholesky decomposition starts with LU-decomposition

- Another property of pd is the possibility of Cholesky decomposition
- Cholesky decomposition starts with your favorite LU-decomposition

$$\overset{\text{pd}}{S} = \begin{bmatrix} 5 & 2 & 0 \\ 2 & 6 & 2 \\ 0 & 2 & 7 \end{bmatrix} \underset{R_2 \leftarrow R_2 - \frac{2}{5}R_1}{\sim} \begin{bmatrix} 5 & 2 & 0 \\ 0 & 26/5 & 2 \\ 0 & 2 & 7 \end{bmatrix} \underset{R_3 \leftarrow R_3 - \frac{5}{13}R_2 \checkmark}{\sim} \begin{bmatrix} 5 & 2 & 0 \\ 0 & 26/5 & 2 \\ 0 & 0 & 81/13 \end{bmatrix} = U$$

and

$$L = \begin{bmatrix} 1 & & \\ 2/5 & 1 & \\ 0 & 10/26 & 1 \end{bmatrix}$$

Thus,

$$S = \overset{L}{\begin{bmatrix} 1 & & \\ 2/5 & 1 & \\ 0 & 10/26 & 1 \end{bmatrix}} \overset{U}{\begin{bmatrix} 5 & 2 & 0 \\ & 26/5 & 2 \\ & & 81/13 \end{bmatrix}}$$

$$= \begin{bmatrix} 1 & & \\ 2/5 & 1 & \\ 0 & 10/26 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{5} & & \\ & \sqrt{26/5} & \\ & & \sqrt{81/13} \end{bmatrix}^2 \begin{bmatrix} 1 & 2/5 & 0 \\ & 1 & 2\times\frac{5}{26} \\ & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & & \\ 2/5 & 1 & \\ 0 & 10/26 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{5} & & \\ & \sqrt{26/5} & \\ & & \sqrt{81/13} \end{bmatrix} \begin{bmatrix} \sqrt{5} & & \\ & \sqrt{26/5} & \\ & & \sqrt{81/13} \end{bmatrix} \begin{bmatrix} 1 & 2/5 & 0 \\ & 1 & 10/26 \\ & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{5} & & \\ 2/\sqrt{5} & \sqrt{26/5} & \\ 0 & \sqrt{20/26} & \sqrt{81/13} \end{bmatrix} \begin{bmatrix} \sqrt{5} & 2/\sqrt{5} & 0 \\ & \sqrt{26/5} & \sqrt{20/26} \\ & & \sqrt{81/13} \end{bmatrix} = LU = LL^T$$

Cholesky decomposition

- $A = LU = LL^t$, i.e. $U = L^t$.
- This is possible when A is pd (symmetric with eigenvalues all positive)
- Some analogy for covariance matrix Σ
  - In univariate setting,
    - σ^2 **vs** σ
    - (variance) **vs** (standard deviation)
  - In multivariate setting,
    - Σ **vs** L (where $\Sigma = LL^t$)
    - (Covariance matrix) **vs** ~~(Standard deviation matrix)~~
    - No terminology such as 'Standard deviation matrix', but L is like a standard deviation in multivariate statistics.
- Applications in simulating normal random variable.
  - In univariate setting,
    - $Z \sim N(0,1) \Rightarrow \mu + \sigma Z \sim N(\mu, \sigma^2)$
  - In multivariate setting,
    - $Z \sim N(0,I) \Rightarrow \mu + LZ \sim N(\mu, \Sigma)$,
    - where I is identity matrix and $\Sigma = LL^t$

Do it yourself

$$S = \begin{bmatrix} 1 & 9 \\ 9 & 100 \end{bmatrix} \sim \begin{bmatrix} 1 & 9 \\ 0 & 19 \end{bmatrix} = U$$

$R_2 \leftarrow R_2 - 9R_1$

LU factorization.

$L = \begin{bmatrix} 1 & 0 \\ 9 & 1 \end{bmatrix}$ $U = \begin{bmatrix} 1 & 9 \\ 0 & 19 \end{bmatrix}$

$\begin{bmatrix} 1 & 0 \\ 9 & 1 \end{bmatrix}\begin{bmatrix} 1 & 9 \\ 0 & 19 \end{bmatrix} = \begin{bmatrix} 1 & 9 \\ 9 & 100 \end{bmatrix}$

Cholesky decomposition.

$$\begin{bmatrix} 1 & 0 \\ 9 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{1} & 0 \\ 0 & \sqrt{19} \end{bmatrix}\begin{bmatrix} \sqrt{1} & 0 \\ 0 & \sqrt{19} \end{bmatrix}\begin{bmatrix} 1 & 9 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 9 & \sqrt{19} \end{bmatrix}\begin{bmatrix} 1 & 9 \\ 0 & \sqrt{19} \end{bmatrix}$$

Check yourself

- Given symmetric matrix, can you perform orthogonal diagonalization?
- If the symmetric matrix is pd (now this is a legit covariance matrix), then can you interpret the results of orthogonal diagonalization as PCA?
- Understands R^2 in geometric sense
- Able to write ordinary and equation.
- Perform Cholesky decomposition to a pd matrix by doing LU and some more treatment afterward?

```
"Optimism is the faith that leads to achievement - Hellen Keller"
```