Quadratic Form and Covariance Matrix

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

서울과학기술대학교
SEOULTECH SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

I. Notice & Review

Understanding θ (L6.p11)

- θ can differ, what does it mean?
  - $\theta = 0^\circ$
    - Lin. Reg. reflects reality ( )
    -
  - Small θ
    - Ax and b are
    - Lin. Reg. reflects reality ( )
  - Large θ
    - Ax and b are
    - Lin. Reg. reflects reality ( )
  - $\theta = 90^\circ$
    - Lin. Reg. reflects reality ( )
    -
- $Cos\ \theta =$
  - [ ] measures explanatory power in percentage term
  - [ ] measures the percentage of variations explained by linear regression
- One can apply cosine law to find R^2 as well by

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

II. Quadratic forms and definite matrix

Motivation

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

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

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

- **Definition.** A symmetric matrix 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

- 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).

$$[x_1 \ x_2] \begin{bmatrix} 1 & 9 \\ 9 & 100 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1x_1^2 + 18x_1x_2 + 100x_2^2$$

Why is the polynomial positive?

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

$$[x_1\ x_2\ x_3]\begin{bmatrix} | & | & | \\ \mathrm{u}_1 & \mathrm{u}_2 & \mathrm{u}_3 \\ | & | & | \end{bmatrix}\begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix}\begin{bmatrix} - & \mathrm{u}_1 & - \\ - & \mathrm{u}_2 & - \\ - & \mathrm{u}_3 & - \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Letting y $= U^t$x gives

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

- 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 $\longrightarrow$ both linear
- Non-Linear Programming
  - Semi-definite programming
    - Objective function & constraints $\longrightarrow$ 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, 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: $\lambda_1 = 9, \lambda_2 = 6, \lambda_3 = 3$,

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

Principal component analysis (PCA)

From the previous example, we have

$$S = PDP^t = \begin{bmatrix} | & | & | \\ \mathrm{u}_1 & \mathrm{u}_2 & \mathrm{u}_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} 9 & & \\ & 6 & \\ & & 3 \end{bmatrix} \begin{bmatrix} - & \mathrm{u}_1 & - \\ - & \mathrm{u}_2 & - \\ - & \mathrm{u}_3 & - \end{bmatrix},$$

where

$$\mathrm{u}_1 = \frac{1}{3}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \ \mathrm{u}_2 = \frac{1}{3}\begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}, \ \mathrm{u}_3 = \frac{1}{3}\begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$$

- When a covariance matrix went throught orthogonal diagonalization, we call $\mathrm{u}_1, \mathrm{u}_2, \mathrm{u}_3$ as **principal components(PC)** of original data.
- 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

$$S = \begin{bmatrix} 5 & 2 & 0 \\ 2 & 6 & 2 \\ 0 & 2 & 7 \end{bmatrix} \sim \begin{bmatrix} 5 & 2 & 0 \\ 0 & 26/5 & 2 \\ 0 & 2 & 7 \end{bmatrix} \sim \begin{bmatrix} 5 & 2 & 0 \\ 0 & 26/5 & 2 \\ 0 & 0 & 81/13 \end{bmatrix} = U$$

and

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

Thus,

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

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

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

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 weighted normal 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"
```