Lecture Note on Principal Component Analysis

This version: September 19, 2022

1 Covariance

Definition 1. *Let* $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ *be a random vector. The covariance matrix of* $\boldsymbol{X}$*, denoted as* $\boldsymbol{\Sigma}$*, is defined as*

$$\boldsymbol{\Sigma}_{p\times p} = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right])(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right])'\right] = \Big(\sigma_{ij}\Big)_{p\times p},$$

where

$$\sigma^2_{i,j} = \begin{cases} Var(X_i), \text{ if } i = j, \\ Cov(X_i, X_j), \text{ if } i \neq j. \end{cases}$$

Example 1. *Let* Z_1 *and* Z_2 *be two independent* $N(0,1)$*. Suppose that* $X_1 = Z_1 + 2Z_2$ *and* $X_2 = Z_2$*. Then*

$$\begin{aligned} Var(X_1) &= Var(Z_1 + 2Z_2) = Var(Z_1) + 4Var(Z_2) = 1 + 4 = 5 \\ Var(X_2) &= Var(Z_2) = 1 \\ Cov(X_1, X_2) &= \mathbb{E}\left[(X_1 - \mathbb{E}\left[X_1\right])(X_2 - \mathbb{E}\left[X_2\right])\right] \\ &= \mathbb{E}\left[X_1 X_2\right] \\ &= \mathbb{E}\left[(Z_1 + 2Z_2)(Z_2)\right] \\ &= \mathbb{E}\left[2Z_2^2\right] && (Z_1, Z_2 \textit{ independent}) \\ &= 2(Var(Z_2) + (\mathbb{E}\left[Z_2\right])^2) = 2 \end{aligned}$$

Hence, the covariance matrix of (X_1, X_2) *is*

$$\mathbf{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}.$$

Remark 1. *The off-diagonal terms of* $\mathbf{\Sigma}$ *are* $Cov(X_i, X_j)$ *for any* $1 \leq i \neq j \leq p$*, so if each pair of components of the random vector* $\boldsymbol{X}$ *is mutually independent, then* $\mathbf{\Sigma}$ *is a diagonal matrix.*

Remark 2. *In matrix notation,*

$$\mathbf{\Sigma} = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])'\right].$$

Remark 3. *Because*

$$\mathbf{\Sigma}' = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])'\right] = \mathbf{\Sigma},$$

$\mathbf{\Sigma}$ *is symmetric. Or, in scalar notation,*

$$\sigma_{i,j} = Cov(X_i, X_j) = Cov(X_j, X_i) = \sigma_{j,i}.$$

2 Principal Component Analysis (PCA)

PCA attempts to summarize variation in the random vector $\boldsymbol{X}$ with few principal components (PC). In below, we will define what are *principal components* and explain how to derive them.

The first PC, PC_1, can be obtained from the following maximization problem:

$$\max_{\boldsymbol{a}_1 \in \mathbb{R}^p} Var(\boldsymbol{a}_1'\boldsymbol{X}) \quad \text{s.t. } \boldsymbol{a}_1'\boldsymbol{a}_1 = 1.$$

The first principal component $PC_1 = \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots a_{1p}X_p$, where $\boldsymbol{a}_1$ is the solution to the above problem. We call $\boldsymbol{a}_1$ the **coefficient** of the first principal component.

The constraint requires $\boldsymbol{a}$ has to be length 1:

$$\boldsymbol{a}'\boldsymbol{a} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1p} \end{pmatrix} = a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1.$$

Remark 4. *The restriction $\boldsymbol{a_i'a_i} = 1$ is necessary, otherwise $Var(\boldsymbol{a}_i'X)$ can be made arbitrary large by multiplying the coefficient with a constant.*

The rest of the principal components can be defined iteratively. [1] The second PC, $PC_2 = \boldsymbol{a}_2'\boldsymbol{x}$, has coefficients $\boldsymbol{a}_2$ given by the optimization problem below:

$$\begin{aligned} \max_{\boldsymbol{a}_2 \in \mathbb{R}^p} \quad & Var(\boldsymbol{a}_2'\boldsymbol{X}) \\ \text{s.t.} \quad & \boldsymbol{a}_2'\boldsymbol{a}_2 = 1, \\ & Cov(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0. \end{aligned}$$

Notice that we require the second PC has to be uncorrelated with the first PC. We can think of it as we are trying to have the second PC explain the variation that is not explained by the first PC.

Similarly, the ith PC, $\boldsymbol{a}_i'\boldsymbol{x}$ is the solution to

$$\begin{aligned} \max_{a_i \in \mathbb{R}^p} \quad & Var(\boldsymbol{a}_i'\boldsymbol{X}) \\ \text{s.t.} \quad & \boldsymbol{a}_i'\boldsymbol{a}_i = 1, \\ & Cov(\boldsymbol{a}_j'\boldsymbol{X}, \boldsymbol{a}_i'\boldsymbol{X}) = 0 \text{ for } j = 1, 2 \dots, i-1. \end{aligned}$$

Remark 5. *Recall that*

$$\boldsymbol{a_i'x} = \langle \boldsymbol{a}, \boldsymbol{x} \rangle = \begin{pmatrix} a_{i1} & a_{i2} & \dots & a_{ip} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p,$$

[1] If the random vector $\boldsymbol{X}$ contains p variables, then there are at most p principal components.

so $\boldsymbol{a}'\boldsymbol{x}$ *is a scalar.*

2.1 Solve PCA when $\boldsymbol{\Sigma}$ Diagonal

Remark 6. *We can verify that*

$$Var(\boldsymbol{a'X}) = \boldsymbol{a'\Sigma a}.$$

See the following example.

Example 2. *Let* $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ *and* $\boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

$$\begin{aligned} Var(\boldsymbol{a'X}) =& Var\left(\begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) \\ =& Var(X_1 + 2X_2) \\ =& Var(X_1) + 4Var(X_2) \\ =& 1 + 4 \cdot 2 = 9 \end{aligned}$$

Alternatively, we can calculate $Var(\boldsymbol{a'X})$ *by*

$$\boldsymbol{a'\Sigma a} = \begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 4 \end{pmatrix} = 9.$$

Indeed, we end up with the same result, as the previous remarks implied.

By remark 6, the optimization problem can be written as

$$\max_{\boldsymbol{a}_1 \in \mathbb{R}^p} \quad \boldsymbol{a_1'\Sigma a_1} \quad \text{s.t. } \boldsymbol{a_1'a_1} = 1.$$

In his book "How to Solve It", mathematician and Probabilist George Pólya famously said

> "If you can't solve a problem, then there is an easier problem you can solve: find it."

Let's follow his suggestion by assuming the covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}_{p\times p},$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. That is, we assume Σ is now a diagonal matrix with positive, decreasing diagonal elements.

Now we are ready to solve the optimization problem for PC_1:

$$\max_{\boldsymbol{a_1}\in\mathbb{R}^p} \quad \boldsymbol{a_1'\Sigma a_1} \quad \text{s.t. } \boldsymbol{a'a} = 1.$$

Notice the objective function is now

$$\begin{aligned}\boldsymbol{a_1'\Sigma a_1} = &\begin{pmatrix} a_{11} & \dots & a_{1p} \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} a_{11} \\ \vdots \\ a_{1p} \end{pmatrix} \\ =&\lambda_1 a_{11}^2 + \lambda_2 a_{12}^2 + \cdots + \lambda_p a_{1p}^2.\end{aligned}$$

By redefining $b_{1i} = a_{1i}^2$, the maximization problem becomes

$$\max_{\boldsymbol{b}} \quad \lambda_1 b_{11} + \cdots + \lambda_p b_{1p} \quad \text{s.t. } b_{1i} + \cdots + b_{1p} = 1.$$

It is plain to see that $b_{11} = 1, b_{12} = \cdots = b_{1p} = 0$ attains the maximum, while subjecting to the constraint. Hence, $a_{11} = \pm 1, a_{12} = \cdots = a_{1p} = 0$ is the solution to the PC_1 problem.[2]

[2] PCs are only uniquely defined up to reflection over the origin. It is not hard to see that if $\boldsymbol{a}^*$ is a solution to the PCA problem, then $-\boldsymbol{a}^*$ is also a solution.

The first PC is hence

$$PC_1 = \boldsymbol{a'X} = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = X_1,$$

which points to the direction with largest variance.

Next we solve the PC_2 problem

$$\begin{aligned} \max_{\boldsymbol{a}_2 \in \mathbb{R}^p} \quad & Var(\boldsymbol{a}_2'\boldsymbol{X}) \\ \text{s.t.} \quad & \boldsymbol{a}_2'\boldsymbol{a}_2 = 1 \\ & Cov(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0 \end{aligned}$$

The constraint

$$\begin{aligned} Cov(\boldsymbol{a_1'X}, \boldsymbol{a_2'X}) =& Cov(X_1, \boldsymbol{a_2'X}) \\ =& \mathbb{E}\left[X_1(a_{21}X_1 + \cdots + a_{2p}X_p)\right] \\ =& a_{21}\mathbb{E}\left[X_1^2\right] = 0, \end{aligned}$$

which implies $a_{21} = 0$. The last equality exploits $\mathbb{E}\left[X_i X_j\right] = 0$ for all $i \neq j$, and $\mathbb{E}\left[X_i\right] = 0$ for all i.

So $a_2 = \begin{pmatrix} 0 \\ a_{22} \\ \vdots \\ a_{2p} \end{pmatrix}$, and the PC_2 problem reduces to

$$\begin{aligned}\boldsymbol{a_2'}\boldsymbol{\Sigma}\boldsymbol{a_2} &= \begin{pmatrix} 0 & a_{22} & \dots & a_{2p} \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} 0 \\ a_{22} \\ \vdots \\ a_{2p} \end{pmatrix} \\ &= \begin{pmatrix} a_{22} & \dots & a_{2p} \end{pmatrix} \begin{pmatrix} \lambda_2 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} a_{22} \\ \vdots \\ a_{2p} \end{pmatrix}.\end{aligned}$$

The problem is now the same as finding PC_1, with one less variable. Hence the solution to the second component is $a_{22} = \pm 1$, $a_{21} = a_{23} = \dots = a_{2p} = 0$, and $PC_2 = \boldsymbol{a_2'}\boldsymbol{X} = X_2$.

Similarly, we have $PC_3 = X_3, PC_4 = X_4, \dots, PC_p = X_p$.

Conclusion: When $\boldsymbol{\Sigma}$ is diagonal, PC_i is X_i, for $i = 1, \dots, p$. So when X_i's are uncorrelated, PCA is basically finding X_i's with largest variances.

2.2 Solve PCA when $\boldsymbol{\Sigma}$ is not Diagonal

What if $\boldsymbol{\Sigma}$ is not diagonal? Luckily, we have the following theorem:

Theorem 1 (Real Spectral Theorem)**.** *If $\boldsymbol{\Sigma}$ is symmetric, then there exists a $p \times p$ matrix $\boldsymbol{P}$ such that*

$$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{\boldsymbol{-1}},$$

where D is a diagonal matrix, and

$$\boldsymbol{P'}\boldsymbol{P} = \boldsymbol{I}_{p\times p},$$

i.e., $\boldsymbol{P}^{-1} = \boldsymbol{P'}$.

Example 3.

$$\begin{pmatrix} 34 & 12 \\ 12 & 41 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{-4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix} \begin{pmatrix} 50 & 0 \\ 0 & 25 \end{pmatrix} \begin{pmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{-4}{5} & \frac{3}{5} \end{pmatrix}.$$

Since $\boldsymbol{\Sigma}$ is symmetric, it can now be written as $\boldsymbol{\Sigma} = \boldsymbol{PDP'}$, where the notation is followed by the Real Spectral theorem. Let $\boldsymbol{b}_i = \boldsymbol{P'a}_i$,

$$\boldsymbol{a_i'\Sigma a_i} = \boldsymbol{a_i'PDP'a_i} = (\boldsymbol{P'a_i})'\boldsymbol{D}(\boldsymbol{P'a_i}) = \boldsymbol{b_i'Db_i}.$$

$$\boldsymbol{a_i'a_i} = \boldsymbol{a_i'(PP^{-1})a_i} = \boldsymbol{a_i'(PP')a_i} = (\boldsymbol{P'a_i})'(\boldsymbol{P'a_i}) = \boldsymbol{b_i'b_i}$$

Restate the maximization problem in terms of $\boldsymbol{b}_i$ gives

$$\max_{\boldsymbol{b}_i \in \mathbb{R}^p} \quad \boldsymbol{b_i'Db_i} \quad \text{s.t. } \boldsymbol{b_i'b_i} = 1,$$

and we know how to solve it since $\boldsymbol{D}$ is diagonal.