Principal Component Analysis

2016160040 최선묵

Dept. of Mathematics
Korea University

Contents

Orthogonal Complement

Orthogonal Complement

Let V be a vector space and let U be its subspace. Then the orthogonal complement of U is the set

$$U^{\perp} = \{\mathbb{v} \in V : \langle \mathbb{u}, \mathbb{v} \rangle = 0, \forall \mathbb{u} \in U\}.$$

Direct Sum

Let U_1 and U_2 be two subspaces of V. For each $\mathbb{v} \in V$, if there exist $\mathbb{u}_1 \in U_1$ and $\mathbb{u}_2 \in U_2$ uniquely such that $\mathbb{v} = \mathbb{u}_1 + \mathbb{u}_2$, then V is the direct sum of U_1 and U_2, and we write $V = U_1 \oplus U_2$.

Properties   If U is a subspace of $\mathbb{R}^n$, then the following hold.

- $\dim(U) + \dim(U^{\perp}) = \mathrm{n}$
- $\mathbb{R}^n = U \oplus U^{\perp}$ $\Leftarrow$ Orthogonal Decomposition
- $(U^{\perp})^{\perp} = U$

Orthogonal Projection and Orthogonal Decomposition

Projection

Let V be a vector space. If $\{\mathbb{u}_1, \ldots, \mathbb{u}_k\}$ is an orthonormal basis for the subspace U of V, then the orthogonal projection of $\mathbb{v} \in V$ onto U is the vector $proj_U \mathbb{v} = \langle \mathbb{v}, \mathbb{u}_1 \rangle \mathbb{u}_1 + \cdots + \langle \mathbb{v}, \mathbb{u}_k \rangle \mathbb{u}_k$.
Generally, a linear map $\pi \colon V \to U$ is a projection if $\pi \circ \pi = \pi$.

Orthogonal Decomposition

Let V be a vector space with $\dim(V) = n$ and let U be its subspace.
Let $\{\mathbb{u}_1, \ldots, \mathbb{u}_k\}$ and $\{\mathbb{u}_{k+1}, \ldots, \mathbb{u}_n\}$ be orthonormal bases for U and $U^\perp$, respectively,
then the orthogonal decomposition of a vector $\mathbb{v} \in V$ is as follows:

$$\mathbb{v} = \sum_{i=1}^{k} a_i \mathbb{u}_i + \sum_{i=k+1}^{n} b_i \mathbb{u}_i \ \in U \oplus U^\perp$$

Here, $a_i = \langle \mathbb{v}, \mathbb{u}_i \rangle$ for $i = 1, \ldots, k$ and $b_i = \langle \mathbb{v}, \mathbb{u}_i \rangle$ for $i = k+1, \ldots, n$.
Note that $\sum_{i=1}^{k} a_i \mathbb{u}_i$ and $\sum_{i=k+1}^{n} b_i \mathbb{u}_i$ are the orthogonal projections onto U and $U^\perp$, respectively.

Covariance

- Variance is defined for a random variable X which tells how far a set of numbers is spread out fro their average value. It is defined mathematically as follows:

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

- Covariance is defined for two random variables X, Y which tells the joint variability. It is defined mathematically as follows:

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

  Note that $Cov(X, X) = Var(X)$ and it has collinearity.

- Covariance matrix Σ can be defined for a random vector $X = (X_1, \ldots, X_n)^\top$ whose elements are as follows:

$$\Sigma_{ij} = \Sigma_{ji} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])], \quad \forall i, j = 1, \ldots, n.$$

  By definition, Σ is symmetric. Note that any covariance matrix is positive semi-definite, that is, $\mathbb{x}^\top \Sigma \mathbb{x} \geq 0$ for any $\mathbb{x} \in \mathbb{R}^n$.

Sample Covariance matrix

- Let $\mathcal{X} = [\chi_1, \ldots, \chi_n]^\top \in \mathbb{R}^{n\times d}$ be a dataset, that is, $\chi_i \in \mathbb{R}^d$ be a sample data.
  Let $\overline{\chi} = (\overline{x}_1, \ldots, \overline{x}_d)$ be a sample mean vector of $\mathcal{X}$ such that $\overline{x}_j = \frac{1}{n}\sum_{i=1}^n (\mathcal{X})_{ij}$.
  Then the sample covariance matrix $\mathcal{S}$ is computed as follows:

$$\mathcal{S} = \frac{1}{n-1}(\mathcal{X} - \overline{\chi})^\top(\mathcal{X} - \overline{\chi}) \in \mathbb{R}^{d\times d}$$

  In short, if we set $Z = \mathcal{X} - \overline{\chi}$, then we have $\mathcal{S} = \frac{1}{n-1}Z^\top Z \in \mathbb{R}^{d\times d}$.
- Note that the denominator is $n-1$ rather than n due to Bessel's correction.
- If the population mean of X_j is known as μ_j for $j = 1, \ldots, d$, then the sample variance is defined as

$$\boldsymbol{S} = \frac{1}{n}(\mathcal{X} - \boldsymbol{\mu})^\top(\mathcal{X} - \boldsymbol{\mu})$$

  where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$.

Principal Components

For any sample covariance matrix $\mathcal{S} \in \mathbb{R}^{d \times d}$,

- it is positive semi-definite, so that its eigenvalues are nonegative.
- it is real symmetric, so that it is orthogonally diagonalizable.

Let $\lambda_1, \ldots, \lambda_d$ be eigenvalues of $\mathcal{S}$ such that $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, and let v_i's be orthonormal eigenvectors corresponding to λ_i's for $i = 1, \ldots, d$.
Then, we call the eigenvectors $v_1, \ldots, v_k$ **the top** k **principal components** of $\mathcal{X}$ for $k \leq d$.

Principal Component Analysis (PCA)

Problem Setup

- Given a dataset $\mathcal{X} = [\chi_1, \ldots, \chi_n]^\top \in \mathbb{R}^{n \times d}$ and a positive integer $k \leq d$, we want to find the best linear projection $\tilde{\mathcal{X}}$ of $\mathcal{X}$ onto a lower dimensional subspace U of $\mathbb{R}^d$ with $\dim(U) = k$. Here, the 'best' linear projection $\tilde{\mathcal{X}} = [\tilde{\chi}_1, \ldots, \tilde{\chi}_n]^\top$ is the linear projection which minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \|\tilde{\chi}_i - \chi_i\|^2.$$

- Sometimes, the compressed/encoded data can be expressed in two ways: $\boldsymbol{z_i} \in \mathbb{R}^k$ and $\tilde{\chi}_i \in \mathbb{R}^d$ which lie in different dimensional spaces. In this sense, $\tilde{\chi}_i \in \mathbb{R}^d$ is called a reconstructed data.
- Note that PCA can be also considered as finding $\tilde{\mathcal{X}}$ which has the smallest reconstruction loss.
- We additionally assume that the population mean of each dimension of data is zero. Then, the sample variance $\mathcal{S}$ of $\mathcal{X}$ will be $\mathcal{S} = \frac{1}{n}\mathcal{X}^\top \mathcal{X} = \frac{1}{n} \sum_{i=1}^{n} \chi_i^\top \chi_i$.

PCA Projection Perspective

Claim: $\tilde{\chi}_i$ is the projection of χ_i onto the subspace spanned by top k principal components of $\mathcal{X}$.

We prove the claim using the mathematical induction on k.

Base case ($k = 1$)

Let v_1 be a normal basis (row) vector of some subspace U of $\mathbb{R}^d$ with $\dim(U) = 1$. Then, we have

$$\tilde{\chi}_i = (v_1 \chi_i^\top) v_1 \quad \text{for all } i = 1, \ldots, n.$$

Then, we need to find $v_1^* = \arg\min_{v_1} \frac{1}{n} \sum_{i=1}^n \| \tilde{\chi}_i - \chi_i \|^2 = \arg\min_{v_1} \frac{1}{n} \sum_{i=1}^n \| (v_1^\top \chi_i) v_1 - \chi_i \|^2$.

$$\begin{aligned} \| \tilde{\chi}_i - \chi_i \|^2 &= (\tilde{\chi}_i - \chi_i)(\tilde{\chi}_i - \chi_i)^\top = (\tilde{\chi}_i - \chi_i)\left(\tilde{\chi}_i^\top - \chi_i^\top\right) = \|\tilde{\chi}_i\|^2 - (\tilde{\chi}_i \chi_i^\top + \chi_i \tilde{\chi}_i^\top) + \|\chi_i\|^2 \\ &= (v_1 \chi_i^\top)^2 \|v_1\|^2 - (v_1 \chi_i^\top)(v_1 \chi_i^\top) - (v_1 \chi_i^\top)(\chi_i v_1^\top) + \|\chi_i\|^2 = \|\chi_i\|^2 - (v_1 \chi_i^\top)^2 \end{aligned}$$

Therefore, it is equivalent to find $v_1^* = \arg\max_{v_1} \frac{1}{n} \sum_{i=1}^n (v_1 \chi_i^\top)^2$.

PCA Projection Perspective

Base case ($k = 1$) continued...

$$\frac{1}{n}\sum_{i=1}^{n}(v_1 x_i^\top)^2 = \frac{1}{n}\sum_{i=1}^{n}(v_1 x_i^\top)(x_i v_1^\top) = v_1\left(\frac{1}{n}\sum_{i=1}^{n} x_i^\top x_i\right) v_1^\top = v_1 \mathcal{S} v_1^\top$$

Now, it is just an optimization problem of finding $v_1^* = \arg\max_{v_1} v_1 \mathcal{S} v_1^\top$ subject to $\|v_1\| = 1$.
We solve the problem by solving Lagrangian $\mathcal{L}(v_1, \lambda_1) = v_1 \mathcal{S} v_1^\top - \lambda_1(v_1 v_1^\top - 1)$.

- $\frac{\partial \mathcal{L}}{\partial \lambda_1} = v_1 v_1^\top - 1 = 0 \quad \Rightarrow \quad v_1 v_1^\top = \|v_1\|^2 = 1.$
- $\frac{\partial \mathcal{L}}{\partial v_1} = 2 v_1 \mathcal{S} - 2\lambda_1 v_1 = 0 \quad \Rightarrow \quad \mathcal{S} v_1^\top = \lambda_1 v_1^\top$ and $v_1 \mathcal{S} v_1^\top = \lambda_1$.

Note that $v_1^\top$ is an eigenvector corresponding to λ_1. Since we are looking for v_1 which maximizes $v_1 \mathcal{S} v_1^\top = \lambda_1$, we can conclude that v_1^* is the eigenvector of $\mathcal{S}$ corresponding to the largest eigenvalue with norm 1. By the definition of the top 1 principal component, the base case of the claim is proved.

PCA Projection Perspective

Induction Hypothesis Assume that the claim holds for $k-1$. ($k \geq 2$)
Choose an ordered orthonormal basis $\{v_1, \dots, v_d\}$ of $\mathbb{R}^d$ and let U be a subspace of $\mathbb{R}^d$ spanned by $\{v_1, \dots, v_{k-1}\}$. Then $U^\perp$ is the subspace spanned by $\{v_k, \dots, v_d\}$. Then, by the orthogonal decomposition, for all $i = 1, \dots, d$, we have

$$\chi_i = \sum_{j=1}^{k-1} a_j v_j + \sum_{j=k}^{d} b_j v_j \in U \oplus U^\perp$$

where $a_j = \chi_i v_j^\top$ and $b_j = \chi_i v_j^\top$. Choose a vector $\varphi \in U^\perp$ such that $\|\varphi\| = 1$. Let V be the subspace spanned by $\{v_1, \dots, v_{k-1}, \varphi\}$, then the projection $\tilde{\chi}_i$ of χ_i onto V is

$$\tilde{\chi}_i = \sum_{j=1}^{k-1} a_j v_j + (\chi_i \varphi^\top)\varphi = \sum_{j=1}^{k-1} (\chi_i v_j^\top) v_j + (\chi_i \varphi^\top)\varphi.$$

Now, we need to find $\{v_1, \dots, v_{k-1}, \varphi\}$ which minimizes $\frac{1}{n} \sum_{i=1}^{n} \|\tilde{\chi}_i - \chi_i\|^2$.

PCA Projection Perspective

Since we have $\|\tilde{\chi}_i - \chi_i\|^2 = \|\chi_i\|^2 - \{\sum_{j=1}^{k-1}(\chi_i v_j^\top)^2 + (\chi_i \varphi^\top)^2\}$, we obtain that

$$\begin{aligned}\{v_1^*, \ldots, v_{k-1}^*, \varphi^*\} &= \arg \min_{v_1,\cdots,v_{k-1},\varphi} \frac{1}{n}\sum_{i=1}^{n} \|\tilde{\chi}_i - \chi_i\|^2 \\ &= \arg \max_{v_1,\cdots,v_{k-1},\varphi} \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k-1}(\chi_i v_j^\top)^2 + \frac{1}{n}\sum_{i=1}^{n}(\chi_i\varphi^\top)^2\right]\end{aligned}$$

Since $v_1, \ldots, v_{k-1}$ and φ are linearly independent, it is equivalent to find $\{v_1^*, \ldots, v_{k-1}^*, \varphi^*\}$ as follows:

$$\{v_1^*, \ldots, v_{k-1}^*\} = \arg \max_{v_1,\ldots,v_{k-1}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k-1}(\chi v_j^\top)^2$$

$$\varphi^* = \arg \max_{\varphi} \frac{1}{n}\sum_{i=1}^{n}(\chi_i\varphi^\top)^2$$

PCA Projection Perspective

By induction hypothesis, we know that $\{v_1^*, \dots, v_{k-1}^*\}$ are the top $k-1$ principal components of $\mathcal{X}$. Therefore, the subspace U is the subspace spanned by top $k-1$ principal components of $\mathcal{X}$.

Moreover, as we did in the base case, we know that $\frac{1}{n}\sum_{i=1}^{n}(\chi_i\varphi^\top)^2 = \varphi\mathcal{S}\varphi^\top$ with $\|\varphi\| = 1$. Again, it is another optimization problem, so we find φ^* by solving Lagrangian $\mathcal{L}(\varphi, \lambda) = \varphi\mathcal{S}\varphi^\top - \lambda(\varphi\varphi^\top - 1)$.

- $\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \quad \Rightarrow \quad \|\varphi\| = 1$
- $\frac{\partial \mathcal{L}}{\partial \varphi} = 2\varphi\mathcal{S} - 2\lambda\varphi = 0 \quad \Rightarrow \quad \mathcal{S}\varphi^\top = \lambda\varphi^\top$ and $\varphi\mathcal{S}\varphi^\top = \lambda$.

Since we are looking for φ which maximizes $\varphi\mathcal{S}\varphi^\top = \lambda$, the vector $\varphi^{*\top}$ should be an eigenvector of $\mathcal{S}$ corresponding to the largest eigenvalue. However, since $\varphi \in U^\perp$ and $U \cap U^\perp = \{0\}$, we know that φ cannot be one of the top $k-1$ principal components, and it should be the top kth principal component of $\mathcal{X}$.

Therefore, the subspace V is the subspace which is spanned by top k principal components of $\mathcal{X}$ and the projection $\tilde{\mathcal{X}}$ of $\mathcal{X}$ onto V minimizes $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\chi}_i - \chi_i\|^2$. Then, by mathematical induction, the claim is proved.

PCA Projection Perspective

From the proof, we have concluded that PCA minimizes the reconstruction loss $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\chi}_i - \chi_i\|^2$. Let's take a look at what the minimized loss will be. Since PCA do the projection, we have that $\chi_i - \tilde{\chi}_i = \sum_{j=k+1}^{d}(\chi_i v_j^\top)v_j$ for all $i = 1, \ldots, n$. Therefore, the reconstruction loss is as follows:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\chi}_i - \chi_i\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=k+1}^{d}(\chi_i v_j^\top)^2 \\
&= \sum_{j=k+1}^{d} v_j\left(\frac{1}{n}\sum_{i=1}^{n}\chi_i^\top \chi_i\right)v_j^\top \\
&= \sum_{j=k+1}^{d}\lambda_j
\end{aligned}$$

PCA Maximum Variance Perspective

<u>Dimension Reduction</u>

- "Retaining most information after data compression is equivalent to capturing the largest amount of variance in the low-dimensional code." (Harold Hotelling, 1933)

Fix $k \leq d$. Let $\tilde{\mathcal{X}} = [\tilde{\chi}_1, \ldots, \tilde{\chi}_n]^\top$ be the 'best' linear projection of $\mathcal{X}$ in the projection perspective. Then, we have $\tilde{\chi}_i = \sum_{j=1}^{k} (\chi_i v_j^\top) v_j$ where v_j is the top jth principal component of $\mathcal{X}$.

<u>Claim</u>: PCA finds the subspace that preserves the largest amount of variance of data in the compressed data. That is, $\tilde{\mathcal{X}}$ maximizes the variance of $\{\chi_1 v_j^\top, \ldots, \chi_n v_j^\top\}$ for all $j = 1, \ldots, k$.

The claim is quite obvious. Since we have assumed that the population mean of each dimension of data is zero, the variance will be the following.

$$\frac{1}{n} \sum_{i=1}^{n} (\chi_i v_j^\top)^2 = \frac{1}{n} \sum_{i=1}^{n} (v_j \chi_i^\top)(\chi_j v_j^\top) = \frac{1}{n} v_j \left(\sum_{i=1}^{n} \chi_i^\top \chi_i \right) v_j^\top = v_j \mathcal{S} v_j^\top$$

We can see that the objective of maximization is exactly the same as in the projection perspective, which proves the claim.

Explained/Captured Variances

Overall, to find an k-dimensional subspace of $\mathbb{R}^d$ that retains as much information as possible, PCA tells us to choose the basis vectors as the k eigenvectors of the sample covariance matrix S that are associated with the k largest eigenvalues.

The maximum amount of variance PCA can capture with the first k principal components is $\sum_{j=1}^{k} \lambda_j$, and the variance lost by data compression via PCA is $\sum_{j=k+1}^{d} \lambda_j$.

Source: Mathematics for Machine Learning (`https://mml-book.github.io/book/mml-book.pdf`)
Based on the lecture MATH292.01 in KU, taught by Donghun Lee (`https://dhl.korea.ac.kr`)