Principal Components and Eigenvectors

date: 20260715
status: #finished
tags: #mathematics #linear_algebra #principal_component_analysis #eigenvector #eigenvalue

---

0. Introduction

(i) Maximum Variance Principle

Given a data matrix $X \in \mathbb{R}^{n \times k}$, we are operate in the k-dimensional subspace spanned by $\{x_i \in \mathbb{R}^k\}_{i=1}^n$ where x_i is the i-th row vector in the data matrix.

Definition (Principle Direction of Variation)

The "principal direction of variation" is simply a unit vector $b_1 \in \mathbb{R}^k$ such that if we do the coordinate projection, getting Xb_1, it will maximize the sample variance **after** projection:

$$b_1 = \max_{b \in \mathbb{R}^n} \frac{1}{n-1} (Xb)' M_n (Xb)$$

Definition (Maximum Variance Principle)

Essentially, we project each x_i onto the one dimensional subspace spanned by b_1, obtaining $b_1' x_i \in \mathbb{R}$. This is so-called dimension reduction but with the "maximum variance principle".

Xb_1 are coordinates of each sample relative to the basis b_1 and $(Xb)' M_n (Xb)$ is the sample variance after projection where $M_n = I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n' \in \mathbb{R}^n$ is a demeaning matrix. We will explain how b_1 can derived in section 1., and have the insight that b_1 is the eigenvector corresponding to the largest eigenvalue of the sample variance-covariance matrix $X' M_n X$.

Note that throughout the note, we are talking about coordinate projection rather than common vector projection.

(ii) Dimension Reduction and Principal Components

From X to Xb_1, we reduce the dimension from k to 1, and a question can be: whether we could reduct to an arbitrary j where $1 < j < k$, applying the same maximum variance principle? The answer is yes, and to do so, we need to construct a matrix $B \in \mathbb{R}^{k\times j}$, and columns of matrix B are the basis of a j-dimensional subspace, and it's so-called **principal components**.

Essentially, we are projecting to the j-dimensional subspace of which columns of B are the basis. We will show in section II. that basis are simply eigenvectors of the sample variance-covariance matrix $X'M_nX$, and can be obtained through singular vector decomposition.

I. Solving Principal Direction of Variation

Starting from the very basic scenario that we what to project the data matrix $X \in \mathbb{R}^{n\times k}$ onto an one dimensional subspace with basis of $b_1 \in \mathbb{R}^k$ where $\|b_1\|^2 = 1$, i.e., b_1 is the orthonormal basis.

b_1 is called as the principal direction of variation and our goal is to maximize the sample variance after projection. For a sample $x_i \in \mathbb{R}^k$, we can obtain its projected coordinate through $b_1'x_i \in \mathbb{R}$ and stack all samples of projected coordinate vertically:

$$[b_1'(x_1 \ \ldots \ x_n)]' = (b_1'X')' = Xb_1 \in \mathbb{R}^n.$$

Therefore, we can construct an optimization problem:

$$b_1 = \max_b \underbrace{\frac{1}{n-1}(Xb)'M_n(Xb)}_{\text{sample variance after projection}}$$

subject to $\|b\|^2 = 1$. That is, the objective function $\mathcal{J}(b)$ is defined as:

$$\mathcal{J}(b) = \frac{1}{n-1}(Xb)'M_n(Xb) = b'\left[\frac{1}{n-1}(X'M_nX)\right]b = b'\Sigma b$$

where we denote the sample variance-covariance matrix of X as $\Sigma \in \mathbb{R}^{k\times k}$.

The Lagrangian is:

$$L(b) = b'\Sigma b - \lambda(b'b - 1)$$

with the FOC:

$$\partial_b L(b_1) = \Sigma b_1 = \frac{\lambda}{2} b_1.$$

Note that the equation of eigenvector appears only if we have the constraint: $\|b\|^2 = 1$.

---

2. Eigenvectors and Principal Components

(i) The eigenvector with the largest eigenvalue is the principal direction of variation.

Through FOC, we know that b_1 is one of the eigenvectors with **unit length** of the sampling variance-covariance matrix $\frac{1}{n-1} X' M_n X$. Denote the corresponding eigenvalue as $\mu = \frac{\lambda}{2}$. However, there are multiple eigenvectors, and which one can maximize the projected coordinate's variance $\mathcal{J}(b)$? Let's see what's the value of our objective $\mathcal{J}(b)$ evaluated at b_1:

$$\mathcal{J}(b_1) = b_1' \Sigma b_1 \overset{\text{by FOC}}{=} b_1' \mu b_1 = \mu \|b_1\|^2 = \mu$$

Therefore, if b_1 is the eigenvector of Σ with the largest eigenvalue, then $\mathcal{J}(b_1)$ is the maximum, and the variance after projection is simply the corresponding eigenvalue. Denote μ_1 as the largest eigenvalue associated to b_1.

Moreover, the SOC requires the matrix:

$$H_L(b_1) = \partial_b^2 L(b_1) = 2\Sigma - \lambda I = 2(\Sigma - \mu_1 I)$$

to be negative definite. Notice that we are not going to check whether for all $h \in \mathbb{R}^k$,

$$h'(\Sigma - \mu_1 I)h < 0.$$

Instead, we only need to ensure that the above condition will hold for all $h \in \mathcal{T}$ where $\mathcal{T} = \{h \in \mathbb{R}^k : b_1' h = 0\}$.

Proof (b_1 is the Maximizer)

To prove $H_L(b_1)$ is negative definite on $\mathcal{T}$, we need to prove for all $h \in \mathcal{T}$,

$$h'(\Sigma - \mu_1 I)h < 0.$$

As Σ is symmetric, by spectrum theorem, there exists a orthonormal basis for $\mathbb{R}^n$, consisting of eigenvectors of Σ: $\{b_1, ..., b_k\}$.Notice that the corresponding eigenvalue are $\mu_1, ..., \mu_k$ in an descending order. Therefore, for any $h \in \mathcal{T}$, we can find $\{c_i\}_{i=2}^k$ such that h is the linear combination of such basis. i.e., $h = \sum_{i=2}^k c_i b_i$. Notice that b_1 is omitted as $b_1' h = 0$

$$
\begin{aligned}
h'(\Sigma - \mu_1 I)h &= \left(\sum_{i=2}^{K} c_i b_i\right)' \Sigma \left(\sum_{i=2}^{k} c_i b_i\right) - \mu_1 \left(\sum_{i=2}^{k} c_i b_i\right)' \left(\sum_{i=2}^{k} c_i b_i\right) \\
&= \left(\sum_{i=2}^{k} c_i b_i\right)' \left(\sum_{i=2}^{k} c_i \Sigma b_i\right) - \mu_1 \sum_{i=2}^{k} c_i^2 b_i' b_i \\
&= \left(\sum_{i=2}^{k} c_i b_i\right)' \left(\sum_{i=2}^{k} c_i \mu_i b_i\right) - \mu_1 \sum_{i=2}^{k} c_i^2 \\
&= \sum_{i=2}^{k} c_i^2 \mu_i b_i' b_i - \sum_{i=2}^{k} \mu_1 c_i^2 \\
&= \sum_{i=2}^{k} c_i^2 (\mu_i - \mu_1) \\
&\leq 0
\end{aligned}
$$

Therefore, if $\mu_1 > \mu_i$ for $i \geq 2$, then there is a unique maximizer b_1 for $\mathcal{J}(b)$.

(ii) Principal components are eigenvectors of sample variance-covariance matrix.

As mentioned in the proof, the collection of k eigenvectors are orthonormal basis for $\mathbb{R}^k$. Therefore, we can use a subset---$B = \left[b_1, b_2, ..., b_j\right] \in \mathbb{R}^{k \times j}$ where $k < j$---of them to construct basis of a lower dimensional coordinate system to facilitate the dimension reduction. These eigenvectors are referred to as the principal components.

Note that b_m's eigenvalue is greater than or equal to b_n's where $m < n$. That is, columns of B is arranged in the descending order respect to each column's corresponding eigenvalue. The new data matrix can be computed through:

$$
\left((B'B)^{-1}B'X'\right)' = XB(B'B)^{-1} = XB \in \mathbb{R}^{n \times j}
$$

where the dimension reduces from k to j. $(B'B)^{-1}B'X'$ is the formula for the coordinate projection and the transpose after is just to follow the convention that we usually stack the observations vertically. Notice that B is a non-square matrix here so $B'B = I$ is not because B is the orthogonal matrix---$B' = B^{-1}$---but because B is semi-orthogonal.

When $k = j$, i.e., we are using the full eigenvectors to construct an orthonormal basis for $\mathbb{R}^k$, the intuition of this equality is change basis with the special case of orthonormal basis, thereby $B^{-1} = B'$. Moreover, changing basis is basically the same thing with coordinate projection but usually use it to refer to the projection that remains the dimension.

---

3. Obtaining Principal Components through SVD

Through the above sections, we know that the dimension reduction from k to j where $1 \leq j \leq k$ can be done by multiplying the data matrix with the matrix $B \in \mathbb{R}^{k \times j}$, whose columns are orthonormal and are a subset of eigenvectors of the sample variance-covariance matrix: $\Sigma = X' M_n X = (M_n X)'(M_n X)$.

Surprisingly, we can obtain the eigenvectors of Σ through SVD where the decomposing matrix is the de-mean matrix $M_n X$, and $\Sigma = (M_n X)'(M_n X) \in \mathbb{R}^{k \times k}$ is simply the right singular matrix (the left one is $(M_n X)(M_n X)' \in \mathbb{R}^{n \times n}$. Suppose:

$$M_n X \overset{\text{SVD}}{=} UMW^T$$

Eigenvectors of Σ is collected in W, and the corresponding eigenvalue is $M'M$. Once we have the k eigenvectors, we can decide how many of them we are going to use. If we just pick one with the largest eigenvalue, then we reduce to dimension 1 , which is the situation we discuss in section 1.

Moreover, we can also construct Σ through M and W , which is facilitated by the spectral theorem, and the reason to do so is that it's more efficient. What's more, since SVD can be apply to any matrix, and due to the precision issue, the symmetric matrix can be non-symmetric with extremely small difference, thereby computing Σ through SVD is robust.

$$\Sigma = W(M'M)W' \overset{\Lambda := M'M}{=} W\Lambda W'.$$