Machine Learning Methods
Principal Component Analysis

By Zhu Xuelin

Email: `xuelin@u.nus.edu`

September 28, 2024

Principal component analysis (PCA) is a powerful statistical tool for dimension reduction, which has been applied widely in many areas.

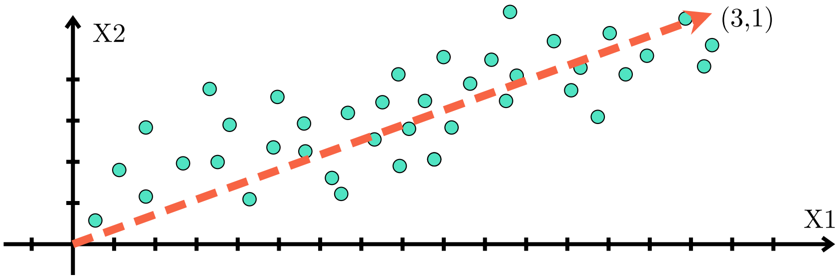
1 Settings and Goals

Suppose we are having such kind of data:

Table 1: Raw data of sample size N and dimension $(q + 4)$

Patient	Response	Age	Gender	Gene 1	Gene 2	$\cdots$	Gene q
#1							
#2							
#3							
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
#N							

The number of column of a data matrix is called the *dimension*. In this case, we are having a data with sample size N and dimension $q+4$. Now we further assume that $N \ll q$, which means the row vectors could not span the $\mathbb{R}^q$ space, and they lie mostly near a low dimensional linear space $\mathbb{R}^p$ $(p \ll q)$. This is the setting of PCA. And here below is an illustration.



As we see, the data do not lie on the whole $\mathbb{R}^2$ space, but nearly concentrate on a line, which is a $\mathbb{R}^1$ linear subspace. And the similar situation may happen in the high dimensional data. Maybe for a 100000-dimensional data, the dimension of the linear space spanned by its row vectors is only 1000. With some approximation, then it could be further reduced to 200-dimension. Our goal with PCA is to find the lower dimensional space $\mathbb{R}^p$, which could characterize the original space $\mathbb{R}^q$.

The most extreme situation is: the n row vectors span a $\mathbb{R}^n$ space. In this case, no lower dimension space could be used for approximation. Therefore, we need to assume that the raw data $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathbb{R}^q$ lie mostly near a low dimension linear space $\mathbb{R}^p$, say $\mathcal{M}$. For this $\mathcal{M}$, there exists an orthogonal base:

$$\boldsymbol{u_1}, \ldots, \mathbf{u}_p \in \mathbb{R}^q \quad \text{such that} \quad \forall \mathbf{v} \in \mathcal{M}, \quad \mathbf{v} = x_1\mathbf{u}_1 + \cdots + x_p\mathbf{u}_p = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_p\end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}.$$

For convenience, we denote

$$U_{q\times p} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_p\end{bmatrix}, \quad \text{with} \quad U^\top U = I_p \ \text{ (note the order!)}.$$

Then every vector in this $\mathcal{M}$ space has a representation using the basis of U's column vectors. Because we assume the raw data $\mathbf{y}$ nearly sits in this space, we have

$$\forall\, \mathbf{y}_i, \quad i \in \{1, \ldots, N\}: \quad \mathbf{y}_i \approx x_1\mathbf{u}_1 + \cdots + x_p\mathbf{u}_p = U\mathbf{x}_i,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is the coordinates under basis U. Now we formally state our assumption of PCA.

> **Problem Setup.** (PCA) Suppose the raw observed data are q-dimensional $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from an unknown population $Y \sim \text{Dist}(\boldsymbol{\mu}, \Sigma)$, and the components of $\mathbf{y}$ nearly lie in a lower p-dimensional linear subspace ($p < q$). Our goal is to find out the p-dimensional subspace, more specifically, to find out the orthogonal basis matrix $U_{q\times p}$ to get
>
> $$\mathbf{y}_{q\times 1} \mapsto \mathbf{x}_{p\times 1} \quad \text{such that} \quad \mathbf{x} = U^\top \mathbf{y} \quad \text{with} \quad U_{q\times p} = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p.\end{pmatrix}$$

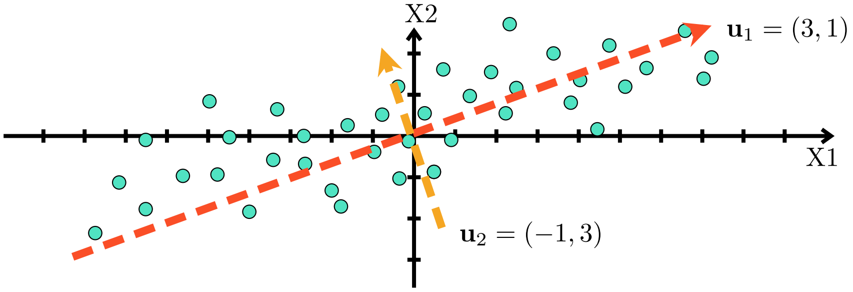
Here diverge two version. If we know the population covariance matrix Σ, then we are doing population PCA, using Σ. However, if we have the sample, we are doing the sample version PCA, using S.

2 Estimation of the Lower Dimensional Space

Now we suppose we are dealing with sample data, instead of the population. There are many ways to estimate the U, one of which was introduced in MA304, using sequential selection by maximization lemma. Now, we introduce another way to derive the first p PCs simultaneously, using SVD or spectral decomposition.

2.1 Case I: $p = 1$

We first deal with the easiest situation: all data nearly lie on a 1 dimensional space. So, our goal is to find the unit direction vector $\mathbf{u}_1$ such that $\mathbf{u}_1^\top \mathbf{u}_1 = 1$. For illustration, let's consider the following example.



If we project the sample on the $\mathbf{u}_1$ direction, the variance of the coordinates will be much larger than that if we project on the $\mathbf{u}_2$ direction. Note that we are projecting only on one direction, so we can express

$$\text{Proj}_{\mathbf{u}_1}(\mathbf{y}_i) = \mathbf{u}_1(\mathbf{u}_1^\top \mathbf{u}_1)^{-1}\mathbf{u}_1^\top \mathbf{y}_i = \mathbf{u}_1\mathbf{u}_1^\top \mathbf{y}_i = \left(\mathbf{u}_1^\top \mathbf{y}_i\right)\mathbf{u}_1 = x_{i1}\mathbf{u}_1,$$

where all x_{i1}'s are scalar, not vector, whose first index i of x_{i1} indicates the obs number, the second 1 indicates the projection on the first direction. If $\mathbf{u}_1$ direction makes the variance of projection coordinates $\{x_{11}, x_{21}, \ldots, x_{N1}\}$ largest, we say $\mathbf{u}_1$ direction captures the most variation of $\mathbf{y}$'s.

Proposition 1. *To find the first PC of the sample $\mathbb{Y}$, we optimize the following function:*

$$\max_{\mathbf{u}_1 \in \mathbb{R}^q} \frac{1}{N-1}\sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2 \quad \textit{subject to} \quad \mathbf{u}_1^\top \mathbf{u}_1 = 1, \quad \textit{where} \quad \bar{x}_1 = \frac{1}{N}\sum_{j=1}^{N} x_{i1}.$$

The solution is

$$\mathbf{u}_1 = \text{the eigenvector of S with the largest eigenvalue.}$$

Proof. It is derived that $x_{i1} = \mathbf{u}_1^\top \mathbf{y}_i$, and then $\bar{x}_1 = \mathbf{u}_1^\top \bar{\mathbf{y}}$. It then follows that

$$\begin{aligned}
\frac{1}{N-1}\sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{u}_1^\top \mathbf{y}_i - \mathbf{u}_1^\top \bar{\mathbf{y}}\right)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\mathbf{u}_1^\top(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top \mathbf{u}_1\right] \\
&= \mathbf{u}_1^\top\left[\frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top\right]\mathbf{u}_1 \\
&= \mathbf{u}_1^\top S \mathbf{u}_1.
\end{aligned}$$

Therefore, our goal becomes to

$$\max_{\mathbf{u}_1 \in \mathbb{R}^q} \mathbf{u}_1^\top S \mathbf{u}_1 \quad \text{subject to} \quad \mathbf{u}_1^\top \mathbf{u}_1 = 1. \tag{1}$$

Introducing a Lagrange multiplier for $\mathbf{u}_1^\top \mathbf{u}_1 = 1$, we set

$$\frac{\partial}{\partial \mathbf{u}_1}\left\{\mathbf{u}_1^\top S \mathbf{u}_1 - \lambda(\mathbf{u}_1^\top \mathbf{u}_1 - 1)\right\}\Bigg|_{\hat{\mathbf{u}}_1} = 0 \quad \Rightarrow \quad S\hat{\mathbf{u}}_1 = \lambda \hat{\mathbf{u}}_1. \tag{2}$$

This means, the optimization is obtained only if $\hat{\mathbf{u}}_1$ is an eigenvector of S. But which eigenvector is the solution? We need to plug (2) back to (1), and see the value of $\mathbf{u}_1^\top S\mathbf{u}_1$.

$$\begin{aligned}\max_{\mathbf{u}_1\in\mathbb{R}^q}\left\{\mathbf{u}_1^\top S\mathbf{u}_1-\lambda(\mathbf{u}_1^\top\mathbf{u}_1-1)\right\}&=\max_{\mathbf{u}_1\in\mathbb{R}^q}\left\{\mathbf{u}_1^\top\lambda\mathbf{u}_1-\lambda\left(\mathbf{u}_1^\top\mathbf{u}_1-1\right)\right\}\\&=\max_{\mathbf{u}_1\in\mathbb{R}^q}\lambda.\end{aligned}$$

By (2), λ is the eigenvalue corresponding to the eigenvector $\mathbf{u}_1$. So, by choosing $\mathbf{u}_1$ to be the eigenvector with the largest eigenvalue, we get this optimized. □

2.2 Case II: $p > 1$

Now we relax the condition $p = 1$, assuming the raw data $\{\mathbf{y}_1, \dots, \mathbf{y}_N\} \in \mathbb{R}^q$ lie nearly on a p dimensional subspace. Suppose $(\mathbf{u}_1, \dots, \mathbf{u}_p)$ is an orthogonal basis of this $\mathbb{R}^p$ space, and we write

$$U=\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p\end{bmatrix}_{q\times p},\quad\text{where}\quad \mathbf{u}_k\in\mathbb{R}^q \text{ for all k} \in\{1,2,\dots,p\}.$$

Then every observation $\mathbf{y}_i$ has p projection coordinates now:

$$\forall i\in\{1,\dots,N\}:\quad \mathbf{y}_i\in\mathbb{R}^q\quad\Longrightarrow\quad \mathbf{x}_i=\begin{pmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{ip}\end{pmatrix}=\begin{pmatrix}\mathbf{u}_1^\top\mathbf{y}_i\\\mathbf{u}_2^\top\mathbf{y}_i\\\vdots\\\mathbf{u}_p^\top\mathbf{y}_i\end{pmatrix}=U^\top\mathbf{y}_i.$$

And as usual, we call $\mathbf{x}_i$ as the projection coordinates of $\mathbf{y}_i$ on the basis U. In fact, we are using the projection to approximate the true $\mathbf{y}_i$, i.e.

$$\mathbf{y}_i\approx x_{i1}\mathbf{u}_1+x_{i2}\mathbf{u}_2+\cdots+x_{ip}\mathbf{u}_p=\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p\end{bmatrix}\begin{pmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{ip}\end{pmatrix}=U_{q\times p}U_{p\times q}^\top\,\mathbf{y}_i.$$

This is an interpretation of what we are doing. Now, let's begin the estimation. Similarly, we also want the projection coordinates capture more variation of $\mathbf{y}$'s.

Proposition 2. *To find the first p PC's of the sample $\mathbb{Y}$, we optimize the following function:*

$$\max_{U\in\mathbb{R}^{q\times p}}\left(\text{sum of sample variance of }\begin{bmatrix}\{x_{11} & x_{21} & \cdots & x_{N1}\}\\\{x_{12} & x_{22} & \cdots & x_{N2}\}\\ & \vdots & & \\\{x_{1p} & x_{2p} & \cdots & x_{Np}\}\end{bmatrix}\right)\quad\textit{subject to}\quad U^\top U=I_p.$$

One solution is $U=\begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_p\end{bmatrix}$, *where*

$$\mathbf{u}_k=\text{the eigenvector of S with the }k\text{−th largest eigenvalue.}$$

And any rotation of the p column vectors $U_{q\times p}O_{p\times p}$, *where* $O^\top O=I_p$, *is another solution.*

Proof. Let's first simplify the optimization function. In Proposition 2, we showed

$$\text{sample variance of } \{x_{11}, \ldots, x_{N1}\} = \mathbf{u}_1^\top S \mathbf{u}_1.$$

So, with the same argument, we have

$$\begin{gathered}\text{sample variance of } \{x_{11}, \ldots, x_{N1}\} = \mathbf{u}_1^\top S \mathbf{u}_1, \\ \vdots \\ \text{sample variance of } \{x_{1p}, \ldots, x_{Np}\} = \mathbf{u}_p^\top S \mathbf{u}_p.\end{gathered}$$

This leads to

$$\text{sum of sample variance of } \begin{bmatrix} \{x_{11} & x_{21} & \cdots & x_{N1}\} \\ \{x_{12} & x_{22} & \cdots & x_{N2}\} \\ & \vdots & & \\ \{x_{1p} & x_{2p} & \cdots & x_{Np}\} \end{bmatrix} = \sum_{i=1}^{p} \mathbf{u}_i^\top S \mathbf{u}_i.$$

Our restrictions for $\mathbf{u}_i$'s are

$$U^\top U = I_p \quad \Longleftrightarrow \quad \begin{cases} \mathbf{u}_i^\top \mathbf{u}_i = 1 & \text{for } j = 1, \ldots, p, \\ \mathbf{u}_i^\top \mathbf{u}_j = 1 & \text{for } i \neq j. \end{cases}$$

Introducing λ_{ij} for every restriction, our optimization function becomes

$$\max_{\mathbf{u}_1, \ldots, \mathbf{u}_p \in \mathbb{R}^q} \left\{ \sum_{i=1}^{p} \mathbf{u}_i^\top S \mathbf{u}_i - \sum_{i=1}^{p} \lambda_{ii} (\mathbf{u}_i^\top \mathbf{u}_i - 1) - \sum_{1 \leq i \neq j \leq p} \lambda_{ij} (\mathbf{u}_i^\top \mathbf{u}_j - 0) \right\}. \tag{3}$$

If we write the Lagrange multipliers as a matrix

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pp} \end{pmatrix},$$

optimization (3) becomes

$$\max_{\mathbf{u}_1, \ldots, \mathbf{u}_p \in \mathbb{R}^q} \Big\{ tr\left(U^\top S U\right) - tr\big[\Lambda \left(U^\top U - I_p\right)\big] \Big\}.$$

Using matrix derivative, and setting the derivative to be zero, we get

$$\frac{\partial}{\partial U} \Big\{ tr\left(U^\top S U\right) - tr\big[\Lambda \left(U^\top U - I_p\right)\big] \Big\} \Big|_{\hat{U}} = \mathbf{0} \quad \Rightarrow \quad S\hat{U} = \hat{U}\Lambda.$$

Note that the Λ matrix is symmetric, so we use spectral decomposition, and it follows that

$$\Lambda = O \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix} O^\top \quad \Rightarrow \quad S\hat{U}O = \hat{U}O \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix}.$$

For convenience, denote $U^* := \hat{U}O$ and suppose $U^* = \begin{bmatrix}\mathbf{u}_1^* & \dots & \mathbf{u}_p^*\end{bmatrix}$. Then we have a very simple expression

$$SU^* = U^* \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix} \quad \Longleftrightarrow \quad \begin{pmatrix} S\mathbf{u}_1^* & S\mathbf{u}_2^* & \cdots & S\mathbf{u}_p^* \end{pmatrix} = \begin{pmatrix} d_1\mathbf{u}_1^* & d_2\mathbf{u}_2^* & \cdots & d_p\mathbf{u}_p^* \end{pmatrix}.$$

This means, the optimization is obtained only if every $\mathbf{u}_k^*$ above is an eigenvector of S, and correspondingly, d_k is the eigenvalue. And here comes the same question: which eigenvectors should we take since S has $\min(N,q)$ eigenvectors? We also plug the possible solution into the maximization function. Before plugging, note that

$$S\hat{U} = \hat{U}\Lambda, \quad U^* = UO \quad \text{and} \quad U^\top U = O^\top (U^*)^\top (U^*) O = I_p,$$

where the last equality holds because $U^* = \begin{bmatrix}\mathbf{u}_1^* & \dots & \mathbf{u}_p^*\end{bmatrix}$ with each $\mathbf{u}_k^*$ being an eigenvectors of S (eigenvectors are orthogonal). Then we look back at the maximization:

$$\begin{aligned}\max_{\mathbf{u}_1,\dots,\mathbf{u}_p \in \mathbb{R}^q} \left\{ tr\left(U^\top S U\right) - tr\left[\Lambda\left(U^\top U - I_p\right)\right] \right\} &= \max_{\mathbf{u}_1,\dots,\mathbf{u}_p \in \mathbb{R}^q} \left\{ tr\left(U^\top U \Lambda\right) - tr\left[\Lambda\left(U^\top U - I_p\right)\right] \right\} \\ &= \max_{\mathbf{u}_1,\dots,\mathbf{u}_p \in \mathbb{R}^q} \left\{ tr\left(\Lambda\right) \right\} \\ &= \max_{\mathbf{u}_1,\dots,\mathbf{u}_p \in \mathbb{R}^q} tr \left\{ O \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix} O^\top \right\} \\ &= \max_{\mathbf{u}_1,\dots,\mathbf{u}_p \in \mathbb{R}^q} \sum_{i=1}^p d_i. \end{aligned}$$

Since every d_k is the eigenvalue corresponding to the eigenvector $\mathbf{u}_k^*$ of S, we choose $d_1, \dots, d_p$ to be the largest eigenvalues, and $\mathbf{u}_k^*$ is the corresponding eigenvectors, it gets optimized. □

Note that the last word of Proposition 2 says: the first p PC's of the sample $\mathbb{Y}$ is NOT unique. The reason for it is we are doing optimization on p PC's simultaneously. This means, **we can rotate** the first p PC's, but remain at the same explained proportion of total variance. However, the explained proportion of each PC from 1 to p may change.