Basic Maths of Principal Component Analysis (PCA)

Ranveer*
CSE, IIT Indore

Principal Component Analysis (PCA) is an extensively used technique in machine learning for reducing the dimensionality and noise of data. It was invented in 1901 by Karl Pearson. In this note, we discuss the basic mathematics behind PCA. We just used basic linear algebra.

1 What are the principal components?

Consider a cricket match between India and New Zealand. Suppose the Indian team is fielding and the New Zealand captain Kane Williamson is at the batting crease, see Figure 1. Which fielder will see Kane's batting stumps best? Consider the views from wicketkeeper Rishab Pant and the point fielder Ravindra Jadeja. Which view is better? Definitely the view from Rishab.

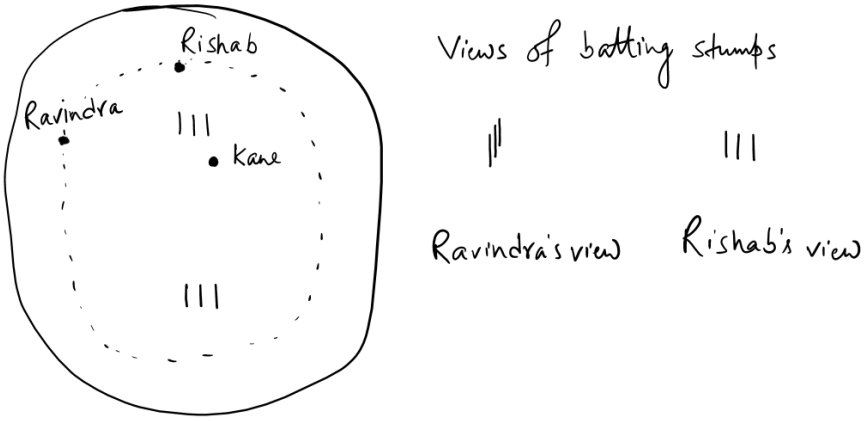
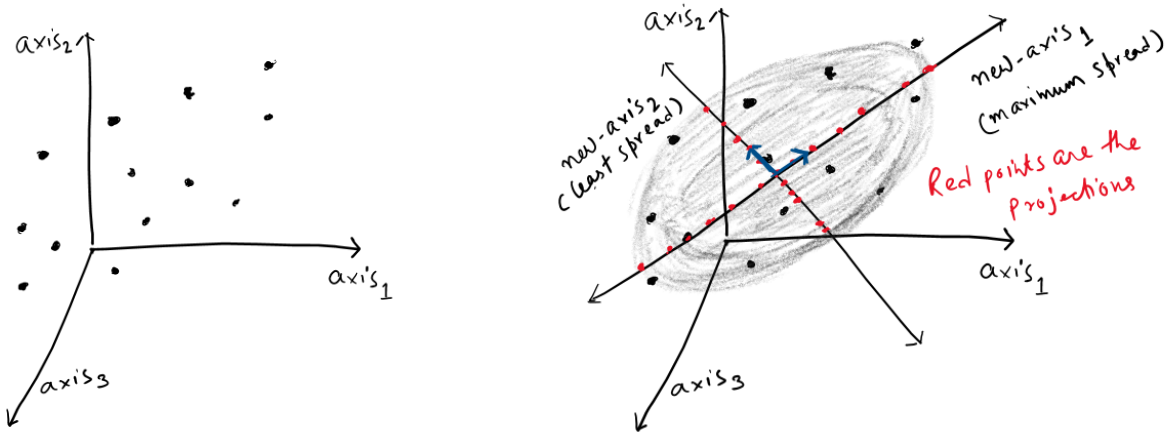


Figure 1

The motive of PCA is similar. For a given set of data points in a d-dimensional space it find the directions along which the data can be seen more clearly. Consider Figure 2.

*Email: ranveer@iiti.ac.in



(a) (b)

Figure 2

In (a) data points are shown in 3-dimensional space. In (b) the same data points can be seen as points in 2-dimensional space (shaded region). Two new orthogonal axes, $\mathbf{new-axis}_1, \mathbf{new-axis}_2$ can measure these points. The red points are the orthogonal projections of the data points on these axes. The $\mathbf{new-axis}_1$ captures the maximum spread of projected data, while $\mathbf{new-axis}_2$ captures the least.

1.1 Basic idea of PCA

Suppose some data points are lying on d-dimensional space. Let $\mathbf{axis}_1, \ldots, \mathbf{axis}_d$ be the orthogonal axes to represent these data point. In order to capture the spread of data, PCA aims at finding new orthogonal axes $\mathbf{new-axis}_1, \ldots, \mathbf{new-axis}_d$ called principal components or principal direction. More precisely, the $\mathbf{new-axis}_1$ captures the maximum spread, $\mathbf{new-axis}_2$ gives the next best maximum spread, and so on. The $\mathbf{new-axis}_d$ captures the least spread, in other words, it captures most of the noise of data. In practice, we need to keep the components that give more spread and discard the ones that give less spread. Thus it is a good technique for reducing the dimensionality and noise in the data.

1.2 Measuring Spread (The Variance)

Consider n 1-dimensional data points $x_1, x_2, \ldots, x_n$. The average of these points is

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}.$$

Their spread is measured by the quantity

$$\text{var}(x) = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n},$$

known as the variance. It measures how the data points are spread with respect to their average.

Let $\mathbf{1}$ denote the all-one vector $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, and $x = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}$. Then the average $\bar{x}$ can be written as

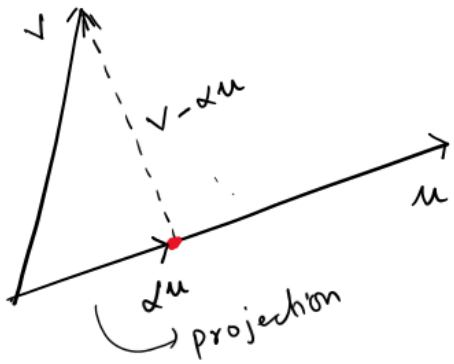
$$\bar{x} = \frac{x\mathbf{1}}{n}. \tag{1}$$

1.3 Projecting a data point onto a line

Consider two d-dimensional vectors

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_d \end{bmatrix}, \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}.$$

Pictorially the projection of v on u is as follows



Note that vectors u and $v - \alpha u$ are orthogonal. So

$$\begin{aligned} u^T(v - \alpha u) &= 0, \\ u^T v - \alpha u^T u &= 0, \\ \alpha &= \frac{u^T v}{u^T u}. \end{aligned}$$

Thus the projection of v on u is the vector

$$\frac{u^T v}{u^T u} u = u^T v \Big(\frac{u}{u^T u} \Big).$$

The unit vector $\frac{u}{u^T u}$ gives the direction. (For all the vector on u the projection of v on these vectors will always be the same.)

2 The first principal component

Given n d-dimensional data points $x_1, x_2 \ldots, x_n$, the data matrix X is follows.

$$X = \begin{bmatrix} | & | & \ldots & | \\ x_1 & x_2 & \ldots & x_n \\ | & | & \ldots & | \end{bmatrix}, \tag{2}$$

that is the i-th column of X is the data point x_i. In machine learning the coordinates of each data x_i are termed as its features. So we can also see X as the matrix where the i-th row is the feature vector f_i, consisting of i-th features of $x_1, x_2, \ldots, x_n$, that is,

$$X = \begin{bmatrix} -f_1- \\ -f_2- \\ \vdots \\ -f_d- \end{bmatrix}. \tag{3}$$

2.1 The aim

Find the unit vector u such that when the data points $x_1, \ldots, x_n$ are projected in the direction of u the variance is maximum. The projected data points are given by the vector

$$u^T \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \ldots & x_n \\ | & | & \cdots & | \end{bmatrix} = u^T X.$$

So the aim is to find u to get

$$\max(\text{var}(u^T X)).$$

The vector u is known as the first principal component.

2.2 Finding the first principal component

We can write

$$\begin{aligned}
\text{var}(u^T X) &= \frac{1}{n}(u^T X - \overline{u^T X}\mathbf{1}^T)(u^T X - \overline{u^T X}\mathbf{1}^T)^T \\
&= \frac{1}{n}\left(u^T X - \frac{u^T X \mathbf{1}}{n}\mathbf{1}^T\right)\left(u^T X - \frac{u^T X \mathbf{1}}{n}\mathbf{1}^T\right)^T \\
&= \frac{1}{n}\left(u^T X - \frac{u^T X \mathbf{1}}{n}\mathbf{1}^T\right)\left(X^T u - \mathbf{1}\frac{\mathbf{1}^T X^T u}{n}\right) \\
&= \frac{1}{n}u^T\left(X - \frac{X\mathbf{1}}{n}\mathbf{1}^T\right)\left(X^T - \mathbf{1}\frac{\mathbf{1}^T X^T}{n}\right)u \\
&= \frac{1}{n}u^T\left(X - \frac{X\mathbf{1}}{n}\mathbf{1}^T\right)\left(X - \frac{X\mathbf{1}}{n}\mathbf{1}^T\right)^T u \\
&= u^T \frac{\tilde{X}\tilde{X}^T}{n} u,
\end{aligned}$$

where

$$\tilde{X} = \begin{bmatrix} -f_1- \\ -f_2- \\ \vdots \\ -f_d- \end{bmatrix} - \begin{bmatrix} -\overline{f_1}- \\ -\overline{f_2}- \\ \vdots \\ -\overline{f_d}- \end{bmatrix}. \tag{4}$$

The matrix $\frac{\tilde{X}\tilde{X}^T}{n}$ is the covariance matrix of the data matrix X, let us denote it by S. We can write

$$\text{var}(u^T X) = u^T S u.$$

Thus our aim is to find u to get

$$\max(u^T S u).$$

(Note that the order of S is $d \times d$).

2.2.1 Finding u

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be the eigenvalues of S, and the corresponding unit eigenvectors be $v_1, \ldots, v_d$. Since S is a symmetric matrix $v_1, \ldots, v_n$ are orthonormal to each other [1]. Write u as a linear combination of $v_1, \ldots, v_d$. That is

$$u = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_d v_d. \tag{5}$$

[1] See Spectral Theorem

We have

$$u^T u = \alpha_1^2 + \cdots + \alpha_d^2 = 1.$$

Now,

$$\begin{aligned} u^T S u &= u^T(\alpha_1\lambda_1 v_1 + \cdots + \alpha_d\lambda_d v_d) \\ &= \lambda_1\alpha_1^2 + \cdots + \lambda_d\alpha_d^2 \\ &\leq \lambda_1(\alpha_1^2 + \cdots + \alpha_d^2) \\ &= \lambda_1. \end{aligned}$$

So we get

$$u^T S u \leq \lambda_1.$$

But since $Sv_1 = \lambda_1 v_1$, this implies

$$v_1^T S v_1 = \lambda_1.$$

Hence

$$\max(u^T S u) = \lambda_1,$$

and this happens when $u = v_1$. Thus the principal component u is the unit eigenvector corresponding to the largest eigenvalue λ_1 of the covariance matrix S. The corresponding variance is λ_1.

3 Other principal components and conclusions

We have seen that the unit eigenvector corresponding to the largest eigenvalue of the covariance matrix S gives the first principal component of X. What about the other principal components? We prove that the i-th principal component is v_i with the corresponding variance equals to λ_i.

We prove this using induction on i. When $i = 1$ we have seen v_1 is the first principal component. Assume that the result is true for an $i \geq 1$, that is $v_1, v_2, \ldots, v_i$ are the first i principal components and the corresponding variances are $\lambda_1, \lambda_2, \ldots, \lambda_i$, respectively. Let u be the $(i+1)$-th principal component. As in 5 write u as a linear combination of $v_1, v_2, \ldots, v_d$. Note that u has to be orthogonal to all of $v_1, v_2, \ldots, v_i$. Hence, by 5

$$\alpha_1 = \alpha_2 = \cdots = \alpha_i = 0.$$

So

$$\begin{aligned} u^T S u &= u^T(\alpha_{i+1}\lambda_{i+1} v_{i+1} + \cdots + \alpha_d\lambda_d v_d) \\ &= \lambda_{i+1}\alpha_{i+1}^2 + \cdots + \lambda_d\alpha_d^2 \\ &\leq \lambda_{i+1}(\alpha_{i+1}^2 + \cdots + \alpha_d^2) \\ &= \lambda_{i+1}. \end{aligned}$$

So we get

$$u^T S u \leq \lambda_{i+1}.$$

But since $Sv_{i+1} = \lambda_{i+1} v_{i+1}$, this implies

$$v_{i+1}^T S v_{i+1} = \lambda_{i+1}.$$

Hence

$$\max(u^T S u) = \lambda_{i+1},$$

and this happens when $u = v_{i+1}$. Thus the $(i+1)$-th principal component is the unit eigenvector corresponding to the $(i+1)$-th largest eigenvalue of the covariance matrix S. The corresponding variance is λ_{i+1}.

Since the important principal components are the ones that give good variance, we ignore the principal components that give less variance as they mainly capture the noise. Typically the number of principal components we consider for our purpose is significantly less than the original data dimension d. Hence PCA is a good technique for reducing the dimensionality and noise in the data.