Chapter 13

Other Kernel Methods K-PCA, K-CCA, K-PLS, K-ICA

13.1 PRINCIPAL COMPONENTS ANALYSIS

Principal component analysis (PCA) is a linear transformation technique that transforms the data to a new coordinate system such that the greatest variance by projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA can be used for dimensionality reduction in a data set while retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring, higher-order ones. Such low-order components often contain the 'most important' aspects of the data. But this is not necessarily the case, depending on the application.

Rotation as a Transform

Let us consider a hypothetical problem of data compression. Assume that we have eight data sets as shown below:

1	1
2	2.5
3	4
4	4.2
-1	-1.5
-2	-2.8
-3	-3.5
-4	-3.9

Now we have a total of 16 data to be transmitted. Can we somehow reduce the number of data to be transmitted if the user of the data can tolerate some amount of error in each data. In this case the answer is yes.

Before going further, let us dwell a bit on the notation. Consider following 8 data points.

	X_1	X_2
$\boldsymbol{x}_1^T$	1	1
$\boldsymbol{x}_2^T$	2	2.5
$\boldsymbol{x}_3^T$	3	4
$\boldsymbol{x}_4^T$	4	4.2
$\boldsymbol{x}_5^T$	−1	−1.5
$\boldsymbol{x}_6^T$	−2	−2.8
$\boldsymbol{x}_7^T$	−3	−3.5
$\boldsymbol{x}_8^T$	−4	−3.9

Note that, looking horizontally we have data points, and looking vertically we have variables. X_1 and X_2, which are assumed to be normalised random variables (here with a mean of zero) . Each data point must be viewed as a point in R^2 . The plot of the data is as shown below.

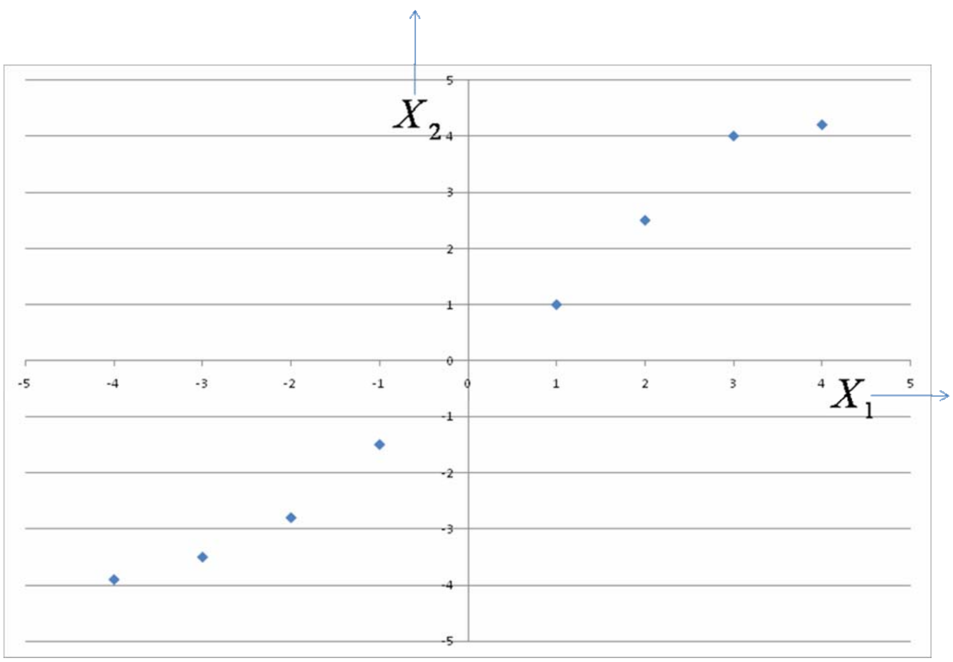

Figure 13.1 Plot of the 2-D data

Note that most of the points are aligned along one axis. Statistician would claim will tell us that most of the variation of the data is along that axis. Now let us choose a unit vector along that axis , that is 45 degree to the x -axis .

Let $\boldsymbol{\beta}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ be the unit vector along the principal axis . Now each point (considered as a vector from origin) is projected on to the axis defined by this unit vector. The distance of such points from the origin is easily computed as $\langle \boldsymbol{x}_i , \boldsymbol{\beta}_1 \rangle = \boldsymbol{x}_i^T \boldsymbol{\beta}_1$. Or the new approximate point obtained by projection is $\tilde{\boldsymbol{x}}_i = \langle \boldsymbol{x}_i , \boldsymbol{\beta}_1 \rangle \boldsymbol{\beta}_1$

The distances of the projected point from the origin are as follows:

$$\begin{array}{c} r_1 \\ 2/\sqrt{2} \\ 4.5/\sqrt{2} \\ 7/\sqrt{2} \\ 8.2/\sqrt{2} \\ -2.5/\sqrt{2} \\ -4.8/\sqrt{2} \\ -6.5/\sqrt{2} \\ -7.9/\sqrt{2} \end{array}$$

These eight values and the angle that $\boldsymbol{\beta}_1$ makes with the x-axis are sent on the transmission channel. The original data can almost be represented. Thus, only 8+ 1=9 numbers need to be sent . The reconstructed values are $\tilde{\boldsymbol{x}}_i = r_{1i}\boldsymbol{\beta}_1$ or $\tilde{\boldsymbol{x}}_i^T = r_{1i}\boldsymbol{\beta}_i^T$. The reconstructed values are the rows of the following matrix:

$$\begin{bmatrix} \boldsymbol{1} & \boldsymbol{1} \\ \boldsymbol{2.25} & \boldsymbol{2.25} \\ \boldsymbol{3.5} & \boldsymbol{3.5} \\ \boldsymbol{4.1} & \boldsymbol{4.1} \\ \boldsymbol{-1.25} & \boldsymbol{-1.25} \\ \boldsymbol{-2.4} & \boldsymbol{-2.4} \\ \boldsymbol{-3.25} & \boldsymbol{-3.25} \\ \boldsymbol{-3.95} & \boldsymbol{-3.95} \end{bmatrix}$$

The reconstructed variables are denoted as $\tilde{X}_1$ and $\tilde{X}_2$. In the above hypothetical example, $\boldsymbol{\beta}_1$ was taken by looking at the plot of the data. If dimension of data is high, one cannot plot it or assess in and see which direction the data has a maximum variation. A formal mathematical approach is therefore has to be adopted for resolving the problem. For doing that we need to borrow some concepts from statistics. So let us do the following computational experiments:

1. Compute the sum of variances $V(X_1)+V(X_2)$ and $V(r_1)$. Observe the difference.

Ans: $V(X_1)=$ 7.5 , $V(X_2)=$ 9.805 , $V(X_1)+V(X_2)=$ 17.305, $V(r_1)$ = 17.1525

2. Choose another unit vector perpendicular to $\boldsymbol{\beta}_1$, say $\boldsymbol{\beta}_2$. Then $\boldsymbol{\beta}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$.

   Find $r_{2i} = \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2 \rangle$ for each point $\boldsymbol{x}_i$. Compute $\langle \boldsymbol{x}_i, \boldsymbol{\beta}_1 \rangle \boldsymbol{\beta}_1^T + \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2 \rangle \boldsymbol{\beta}_2^T$ and observe that the result is $\boldsymbol{x}_i^T$.

$$r_2 = \begin{bmatrix} 0 \\ -0.5/\sqrt{2} \\ -1/\sqrt{2} \\ -0.2/\sqrt{2} \\ 0.5/\sqrt{2} \\ 0.8/\sqrt{2} \\ 0.5/\sqrt{2} \\ -0.1/\sqrt{2} \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 2.25 & 2.25 \\ 3.5 & 3.5 \\ 4.1 & 4.1 \\ -1.25 & -1.25 \\ -2.4 & -2.4 \\ -3.25 & -3.25 \\ -3.95 & -3.95 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -0.25 & 0.25 \\ -0.5 & 0.5 \\ -0.1 & 0.1 \\ 0.25 & -0.25 \\ 0.4 & -0.4 \\ 0.25 & -0.25 \\ -0.05 & 0.05 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 2.5 \\ 3 & 4 \\ 4 & 4.2 \\ -1 & -1.5 \\ -2 & -2.8 \\ -3 & -3.5 \\ -4 & -3.9 \end{bmatrix}$$

that is $\langle \boldsymbol{x}_i, \boldsymbol{\beta}_1 \rangle \boldsymbol{\beta}_1^T + \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2 \rangle \boldsymbol{\beta}_2^T = \boldsymbol{x}_i^T$

3. Compute $V(r_2)$.

Ans. $V(r_2)$ = 0.1525

4. Check whether $V(X_1) + V(X_2)$ is equal to $V(r_1) + V(r_2)$

Ans . $V(X_1) + V(X_2)$ = 17.305, $V(r_1) + V(r_2)$ = 17.1525+0.1525=17.305.

5. Show that $\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{x}_i$ gives total variance $V(X_1) + V(X_2)$.

$$\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{x}_i = \frac{1}{m}\left[\boldsymbol{x}_1^T \boldsymbol{x}_1 + \boldsymbol{x}_2^T \boldsymbol{x}_2 + \dots + \boldsymbol{x}_m^T \boldsymbol{x}_m \right]$$

$$= \frac{1}{8}\left[(1^2 + 1^2) + (2^2 + 2.5^2) + \dots + ((-4)^2 + (-3.9)^2) \right] = 17.305$$

Since most of the variance is along $\boldsymbol{\beta}_1$, it is said that the data, for all practical purposes lies within one-dimensional sub-space spanned by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_1$ is called the principal direction. The projections of the data in this direction are the principal components. That is , principal direction is $\boldsymbol{\beta}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ and principal components are given by elements of $\boldsymbol{r}_1$.

PCA Linear Sub-spaces