Principal Component Analysis II

In this section we show how principal components are calculated.

Preliminary Information

For the data set shown below, we give examples of the techniques used in PCA.

x_1	122	21	105	101	155	131	115	53	75	45
x_2	117	32	140	105	149	146	82	60	82	37

Variance and Standard Deviation

The variance, s^2, of a data sample measures the spread of the data, it is defined to be:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

The standard deviation also measures the spread of the data, it is defined to be the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

For the data given shown above:

- the variance of the x_1 variable $s^2_{x_1} = 1803.12$
- the standard deviation of the x_1 variable, $s_{x_1} = 42.46$
- the variance of the x_2 variable $s^2_{x_2} = 1891.33$
- the standard deviation of the x_2 variable, $s_{x_2} = 43.49$

Covariance

The covariance of the two variables x and y is a measure of how much the two variables change together, it is defined to be:

$$Cov\left(x_1, x_2\right) = \frac{\sum_{i=1}^{n}\left(x_{1i} - \overline{x}_1\right)\left(x_{2i} - \overline{x}_2\right)}{n-1}$$

For the data given shown above, the covariance of the variables x and y is $Cov(x_1, x_2) = 1690.22$.The positive value indicates that both variables increase together.

Covariance Matrix and Correlation Matrix

Covariance is always measured between two variables. The covariance matrix for two variables is:

$$\mathbf{S} = \begin{bmatrix} Var\left(x_1\right) & Cov\left(x_1, x_2\right) \\ Cov\left(x_2, x_1\right) & Var\left(x_2\right) \end{bmatrix}$$

Note that $Cov(x_1, x_2) = Cov\left(x_2, x_1\right)$. If we have a data set with more than two variables, there is more than one covariance measurement that can be calculated. For example, the covariance matrix for a data set with three variables, x, x_2, x_3 is

$$\mathbf{S} = \begin{bmatrix} Var\left(x_1\right) & Cov\left(x_1, x_2\right) & Cov\left(x_1, x_3\right) \\ Cov\left(x_2, x_1\right) & Var\left(x_2\right) & Cov\left(x_2, x_3\right) \\ Cov\left(x_3, x_1\right) & Cov\left(x_3, x_2\right) & Var\left(x_3\right) \end{bmatrix}$$

Since

- $Cov\left(x, y\right) = Cov\left(y, x\right)$
- $Cov\left(x, z\right) = Cov\left(z, x\right)$
- $Cov\left(y, z\right) = Cov\left(z, y\right)$

The matrix is symmetrical about the main diagonal.

The covariance matrix for variables x and y is

$$\mathbf{S} = \begin{bmatrix} 1803.12 & 1690.222 \\ 1690.222 & 1891.33 \end{bmatrix}$$

The **correlation matrix** is a scaled version of the covariance matrix. Instead of covariances, the elements of the correlation matrix consist of correlations between two variables given by:

$$Cor\,(x,y) = \frac{Cov\,(x,y)}{s_x s_y}$$

The correlation for the two variables, x and y is given by:

$$Cor\,(x,y) = \frac{1690.222}{42.46 \times 43.49} = 0.915$$

thus, the correlation matrix for the two variables, x and y is given by:

$$\mathbf{S} = \begin{bmatrix} 1 & 0.915 \\ 0.915 & 1 \end{bmatrix}$$

Eigenvalues and Eigenvectors

Given a square matrix, A, a vector v, is said to be an eigenvector of the matrix A if it does not change direction when multiplied by A,i.e.

$$Av = \lambda v$$

where λ is a constant, known as an eigenvalue of A.

Example 1 *Let* $A = \begin{pmatrix} 6 & 9 \\ 6 & 3 \end{pmatrix}$, $v_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ *and* $v_2 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ *then* v_1 *is not an eigenvector of* A *but* v_2 *is an eigenvector of* A *since:*

$$\begin{pmatrix} 6 & 9 \\ 6 & 3 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 21 \\ 15 \end{pmatrix}$$

$$\begin{pmatrix} 6 & 9 \\ 6 & 3 \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 36 \\ 24 \end{pmatrix} = 12\begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

We see that $Av_2 = \lambda v_2$ *where the eigenvalue* $\lambda = 12$.

Euclidean Space

In principal component analysis, the distance between two points in $k-$dimensional space is measured using Euclidean distance. For two points i and j, with $k-$dimensional coordinate values, $\mathbf{x}_i = (x_{i1}, x_{i2}, ...x_{ik})$ and $\mathbf{x}_j = (x_{j1}, x_{j2}, ...x_{jk})$, the Euclidean distance is defined as

$$\begin{aligned} \mathbf{d}_{ij} &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ik} - x_{jk})^2} \\ \mathbf{d}_{ij} &= \sqrt{\sum_{q=1}^{k} (x_{iq} - x_{jq})^2} \end{aligned}$$

We are familiar with this definition in two dimensions (pythagorus). $\mathbf{x}_i = (x_{i1}, x_{i2})$ and $\mathbf{x}_j = (x_{j1}, x_{j2})$

$$\mathbf{d}_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

Principal Component Analysis

The aim of principal component analysis is to represent a k-dimensional data set in m dimensions, where $m < k$ (i.e. dimension reduction). This $m-$dimensional reduction represents the original $k-$dimensional data in the best possible way, by minimising the difference between the representation in $k-$dimensions and in $m-$dimensions. Let $\mathbf{d}_{ij}$ represent the Euclidean distance between points i and j in $k-$ dimensions and let $\widehat{\mathbf{d}_{ij}}$ represent the Euclidean distance between points i and j in m-dimensions. PCA analysis finds the m components such that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{d}_{ij}^2 - \widehat{\mathbf{d}_{ij}}^2\right)$$

is minimised with respect to $\widehat{\mathbf{d}_{ij}}^2$.

Note that $\mathbf{d}_{ij}^2 = \sum_{q=1}^{k}\left(x_{iq} - x_{jq}\right)^2$ and the $\widehat{\mathbf{d}_{ij}}^2$ that minimises the sum above can be written as $\widehat{\mathbf{d}}_{ij}^2 = \sum_{q=1}^{m}\left(p_{iq} - p_{jq}\right)^2$.

Recall, the first principal component is a linear combination of the original variables and is chosen to capture as much of the variation present in the original variables as possible. The second principal component is defined to be a linear combination of the original variables that captures as much of the remaining variation subject to being uncorrelated with (perpendicular to) the first principal component. Subsequent components are defined similarly. We now show how the coefficients specifying the linear combination of the original variables for each component are found. We can write the first principal component $\mathbf{p}_1$ as the linear combination

$$\mathbf{p}_1 = a_{11}x_1 + a_{12}x_2 + ...a_{1k}x_k$$

whose sample variance is the greatest among all such linear combinations. In vector notation, we can write the first principal component as $\mathbf{p}_1 = \mathbf{a}_1^T\mathbf{x}$.

[Aside: Since the variance of $\mathbf{p}_1$ could be increased without limit by simply increasing the coefficients $\mathbf{a}_1^T = (a_{11}, a_{12}, ..., a_{1k})$, a restriction must be placed on these coefficients. As we shall see later, a sensible constraint is to require that the sum of squares of the coefficients, $\mathbf{a}_1^T\mathbf{a}_1$, should take the value one.]

To find the coefficients defining the first principal component we need to choose the elements of the vector $\mathbf{a}_1$ so as to maximise the variance of $\mathbf{p}_1$ subject to the constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$.

It can be shown that $\mathbf{a}_1$ is the eigenvector of the sample covariance matrix, $\mathbf{S}$, corresponding to its largest eigenvalue.

The second principal component $\mathbf{p}_2 = \mathbf{a}_2^T\mathbf{x}$ is the linear combination with the greatest variance subject to the two conditions $\mathbf{a}_2^T\mathbf{a}_2 = 1$ and $\mathbf{a}_2\mathbf{a}_1 = 0$. The second condition ensures that $\mathbf{p}_1$ and $\mathbf{p}_2$ are uncorrelated. The vector of coefficients $\mathbf{a}_2$ that maximises the variance of $\mathbf{p}_2$ subject to the constraints $\mathbf{a}_2^T\mathbf{a}_2 = 1$ and $\mathbf{a}_2\mathbf{a}_1 = 0$ **is the eigenvector of the sample covariance matrix corresponding to the second largest eigenvalue**, λ_2. Similarly, the j^{th} principal component is the linear combination $\mathbf{p}_j = \mathbf{a}_j^T\mathbf{x}$ which has the greatest variance subject to the conditions $\mathbf{a}_j^T\mathbf{a}_j = 1$ and $\mathbf{a}_j\mathbf{a}_i = 0$ (for $i < j$) and it can be shown that $\mathbf{a}_j$ **is the eigenvector of the sample covariance matrix corresponding to the j^{th} largest eigenvalue,** λ_j.

Calculating the Scores

The scores give the coordinates of the observations in the newly defined PC plane.
Let $\mathbf{P}$ represent the **loadings** matrix which consists of the newly defined principal components $(\mathbf{p}_1, \mathbf{p}_2, ..\mathbf{p}_k)$
N.B. we may choose not to retain all k components in our PCA model. If we choose to retain $m < k$ components then $\mathbf{P}$ will have dimensions $m \times k$.
Let $\mathbf{T}$ represent the matrix containing the scores, and let $\mathbf{X}$ represent the scaled data. The scores $\mathbf{T}$ are defined to be:

$$\mathbf{T} = \mathbf{XP}$$

Variation accounted for by the PCA model

The total variance of the k principal components will equal the total variance of the original variables so that

$$\sum_{j=1}^{k} \lambda_j = s_1^2 + s_2^2 + ... + s_k^2$$

where s_j^2 is the sample variance of x_j. For standardised data, we can write this more concisely as:

$$\sum_{j=1}^{k} \lambda_j = trace\,(\mathbf{S})$$

where $\mathbf{S}$ is the correlation matrix. Consequently, the j^{th} principal component accounts for a proportion of the total variation of the original data, calculated as

$$\frac{\lambda_j}{trace\,(\mathbf{S})}$$

The first m principal components, where $m < k$, account for a proportion

$$= \frac{\sum_{j=1}^{m} \lambda_j}{trace\,(\mathbf{S})}$$

We can analyse the data below using PCA

x_1	122	21	105	101	155	131	115	53	75	45
x_2	117	32	140	105	149	146	82	60	82	37

The standardised data is shown in the table below and in Fig 1.

x_1	0.70	-1.68	0.30	0.21	1.48	0.91	0.54	-0.93	-0.41	-1.11
x_2	0.51	-1.44	1.03	0.23	1.24	1.17	-0.30	-0.80	-0.30	-1.33

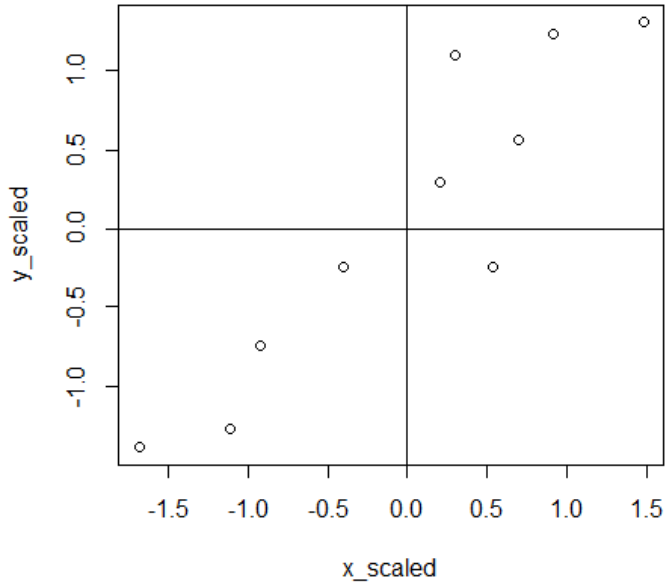

Figure 1

The correlation matrix for the two variables is

$$\mathbf{S} = \begin{bmatrix} 1 & 0.915 \\ 0.915 & 1 \end{bmatrix}$$

the eigenvalues and unit eigenvectors for the matrix are:

$$\lambda_1 = 1.915,\ v_1 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \text{ and } \lambda_2 = 0.085,\ v_2 = \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix}$$

Let's relate these eigenvalues and eigenvectors back to the data. Figure 2 shows a plot of the scaled and centred data with the lines corresponding to the two eigenvectors drawn. The first principal component is given by the eigenvector v_1 which has the largest eigenvalue, we see that this line captures the maximum amount of variation in the data. The second principal component is given by the eigenvector v_2 which is orthogonal to the first principal component.

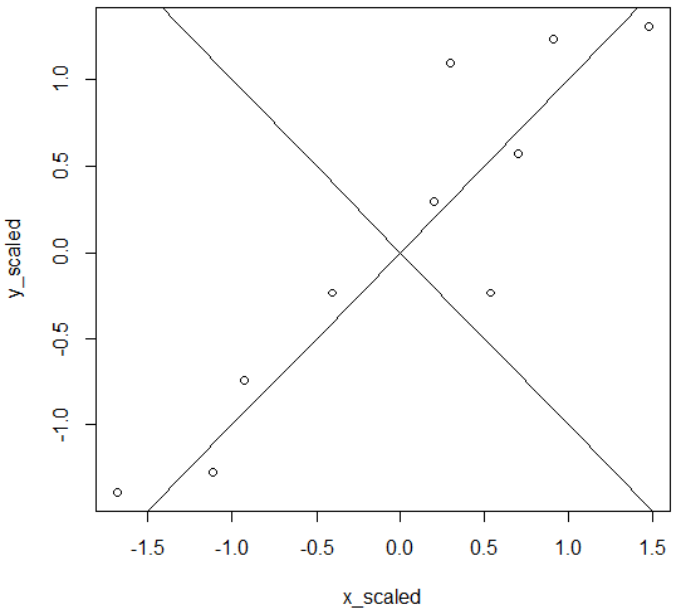

Figure 2: A plot of the scaled and centred data with the lines corresponding to the two eigenvectors representing PC1 and PC2.

Recall that to view the original data in relation to the newly defined principal components we project the original data onto the principal components to obtain the scores which can be plotted in a score plot see Fig 3.

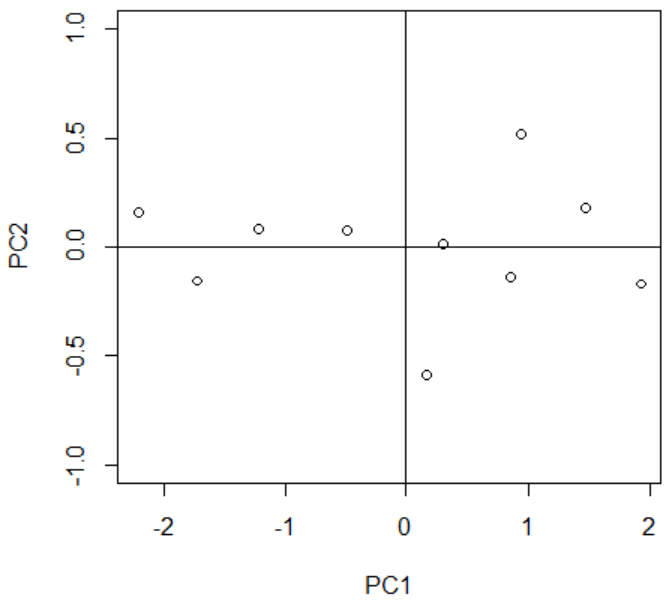

Figure 3

The score plot is obtained by simply rotating the PC lines in Fig. 2.
The first principal component accounts for a proportion of the total variation of the original data, given by

$$\begin{aligned} & \frac{\lambda_1}{trace\,(\mathbf{S})} \\ = & \frac{1.915}{(1.915+0.085)} \\ = & 0.9575 \end{aligned}$$

Recall, the scores, $\mathbf{T}$ are given by

$$\mathbf{T} = \mathbf{XP}$$

The scores depend on how many principal components we retain in our model. We will calculate the scores from a model containing just the first principal component and from the full model retaining both principal components.

In matrix form, the scaled data set is

$$\mathbf{X} = \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix}$$

In matrix form, the loadings are

$$\mathbf{P} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$$

One principal component

If we retain PC1 only then

$$\mathbf{P} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$$

so that

$$\begin{aligned}
\mathbf{T} &= \mathbf{XP} \\
\mathbf{T} &= \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \\
\mathbf{T} &= \begin{bmatrix} 0.85 \\ -2.21 \\ 0.94 \\ 0.31 \\ 1.92 \\ 1.47 \\ 0.16 \\ -1.22 \\ -0.50 \\ -1.73 \end{bmatrix}
\end{aligned}$$

Two principal components

If we retain both principal components then

$$\mathbf{P} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$$

so that

$$
\begin{aligned}
T &= XP \\
T &= \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix} \\
T &= \begin{bmatrix} 0.85 & -0.14 \\ -2.21 & 0.16 \\ 0.94 & 0.52 \\ 0.31 & 0.02 \\ 1.92 & -0.16 \\ 1.47 & 0.18 \\ 0.16 & 0.59 \\ -1.22 & 0.09 \\ -0.50 & 0.08 \\ -1.73 & -0.16 \end{bmatrix}
\end{aligned}
$$

We can compare the values in the matrix with the score plot above.