A tutorial on Principal components analysis

This tutorial aims to improve the understanding of principal component analysis (PCA) for readers that are barely familiar with this technique. The following topics are covered regarding PCA: Aim, applications, assumptions, estimation, visualization, data preprocessing and outliers detection.

Aim of PCA

In a data matrix $\mathbf{X}_{IxJ}$ with *I* samples and *J* variables, PCA aims to find linear combinations of the variables for which the samples show the largest variation. PCA will extract a total of *J* linear combinations. The first linear combination will correspond to the direction in which the samples show the largest variation, but respecting some PCA assumptions. The second linear combination will correspond to the second direction in which the samples show the largest variation, also respecting the assumptions. And so on until the last PC will correspond to the direction in which the data shows the smallest variation respecting the assumptions. These linear combinations, often called principal components (PCs), can be regarded as a new set of variables in which the data can be plotted.

Applications of PCA

The most common application of PCA is to increase the interpretability of the data. Since the data shows large variation for the first PCs, it is often possible to achieve a fair representation of the data with just a few of the first PCs. Another application of PCA is to detect noise in the data. PCA achieves this by assuming that the data has a high signal to noise ratio, and therefore the signal of the data is modeled in the first PCs. The noise, on the contrary, is modeled in the last PCs (PCs for which the data shows very little variation).

Assumptions

In PCA, the PCs are extracted one after another according to the aim (maximizing the variance of the data), and under two main assumptions:

1) The PCs need to be orthogonal to each other. This enables to separate signal from noise, since in statistics the residuals (noise) should be independent of the signal.
2) For each PC, the sum of the squares of the coefficients that relate the PC to the original variables need to sum up to 1. This enables to preserve the distance between samples. In other words, the Euclidean distance between two samples will be the same in the spaces defined by the original variables and by the PCs.

Assumptions. Mathematical notation and examples

We will now express these assumptions in mathematical notation. Let's assume for now that the data matrix $\mathbf{X}_{ixj}$ has been mean centered and standardized to variance 1 for the variables (in columns). We will return to this in the section of data preprocessing. To illustrate the explanation we will use the Iris data that is integrated in the software R. This dataset consists of 150 samples and 4 quantitative variables.

Let's first express in mathematical notation that the PCs are linear combinations of the original variables. In mathematical notation this can be written as:

$$\mathbf{t}_1=\mathbf{x}_1\, p_{11}+\mathbf{x}_2\, p_{12}+\ldots+\mathbf{x}_j p_{1j} = \mathbf{X}\, \mathbf{p}_1 \qquad \textit{Equation 1}$$

Where $\mathbf{t}_1$ is a column vector with *i* rows containing the coordinates of the samples on the first PC (PC$_1$), $\mathbf{x}_1$,$\mathbf{x}_2$...$\mathbf{x}_j$ are column vectors containing the coordinates of the samples in each of the *j* original variables, and p_{11},p_{12}...p_{1j} are the coefficients that relate PC$_1$ to the *j* original variables, $\mathbf{X}_{ixj}$ is the data matrix containing the samples in rows and the original variables in columns, and $\mathbf{p}_1$ is a column vector with *j* rows containing the coefficients p_{11},p_{12}...$p_{1j.}$

Let’s assume for now that we know how to estimate PC_1. After centering the data, $\mathbf{p}_1$ for the Iris data is as shown in Table 1:

Table 1. PC_1 in Iris data

	$\mathbf{p}_1$
Sepal.Length	0.53
Sepal.Width	-0.27
Petal.Length	0.59
Petal.Width	0.57

In Table 1 we observe that $\mathbf{p}_1$ is a collection of coefficients for the original variables. These coefficients enable to transform the space defined by the original variables into another space defined by PC_1. Thus, these coefficients allow to create a new axis, and at the same time they define the new axis. For this reason, we say that each PC is both, a set of coefficients and a direction. For the Iris data, the new axis is geometrically defined by connecting the 4-dimensionals dot in Table 1 with the 4-dimensions-origin.

If we wish to obtain the coordinates of sample 1 on PC_1 we would just need to use these coefficients. As an example, Table 2 shows how the coordinates of PC_1 are computed for the first sample of the Iris data.

Table 2. Coordinates of sample 1 in the original space and in the PC space.

	Sepal.Length	**Sepal.Width**	**Petal.Length**	**Petal.Width**	**PC1 Coordinates**
Coordinates in the original space	5.1	3.5	1.4	0.2	
Computing PC1 coordinates	[5.1*0.53]	+[3.5*(-0.27)]	+[1.4*0.59]	+[0.2*0.57]	2.65

Equation 1 can be extended to *j* PCs.

$$\mathbf{T}=\mathbf{XP} \qquad \textit{Equation 2}$$

Where $\mathbf{T}_{\mathbf{ixj}}$ is a matrix with the coordinates of the samples in the space defined by the PCs, where the samples are in rows and the PCs are in columns, $\mathbf{X}_{\mathbf{ixj}}$ is the data matrix, $\mathbf{P}_{\mathbf{jxj}}$ is a matrix containing the coefficients that relate the PCs to the original variables. $\mathbf{T}_{\mathbf{ixj}}$ is often called the “scores matrix” and $\mathbf{P}_{\mathbf{jxj}}$ is called the “loadings matrix”. For the Iris data, one can think of $\mathbf{P}_{\mathbf{jxj}}$ as an extension of Table 1, in which another three columns have been added (one for each of PC_2, PC_3 and PC_4).

Let’s now express mathematically the two assumptions.

1) The PCs need to be orthogonal to each other.

In mathematical notation this can be expressed as:

$$\mathbf{PP}^{-1}=\mathbf{I} \qquad \textit{Equation 3}$$

Where $\mathbf{I}_{\mathbf{jxj}}$ is the identity matrix. Let’s consider for now that we know the four PCs for the Iris data. Thus, we know $\mathbf{P}_{\mathbf{jxj}}$. $\mathbf{P}$ is a 4x4 matrix, since there are 4 quantitative variables, and therefore $\mathbf{I}$ is a 4x4 Matrix. Equation 3 becomes:

Formula 1. The PCs are orthogonal to each other, Iris data

$$\begin{bmatrix} 0.53 & -0.38 & 0.72 & 0.26 \\ -0.27 & -0.92 & -0.24 & -0.12 \\ 0.59 & 0.02 & -0.14 & 0.8 \\ 0.57 & 0.07 & -0.63 & 0.52 \end{bmatrix} \begin{bmatrix} 0.53 & -0.27 & 0.59 & 0.57 \\ -0.38 & -0.92 & 0.02 & 0.07 \\ 0.72 & -0.24 & -0.14 & -0.63 \\ 0.26 & -0.12 & 0.8 & 0.52 \end{bmatrix} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

Note that the first column of $\mathbf{P}$ in Formula 1 corresponds to the coefficients in Table 1.

2) For each PC, the sum of the squares of the coefficients that relate the PC to the original variables need to sum up to 1.

In mathematical notation this can be expressed as:

$$\mathbf{P}=\mathbf{P}^{(\mathbf{norm})}\mathbf{I} \qquad \textit{Equation 4}$$

Where $\mathbf{P}_{\mathbf{jxj}}$ is the loadings matrix, and $\mathbf{P}^{(\mathbf{norm})}{}_{\mathbf{jxj}}$ is the loadings matrix after dividing each column by the sum of squares, and $\mathbf{I}_{\mathbf{kxk}}$ is the identity matrix.

Figure 1 shows that for the Iris data the sum of the squares of the coefficients sums up to 1 for PC_1.

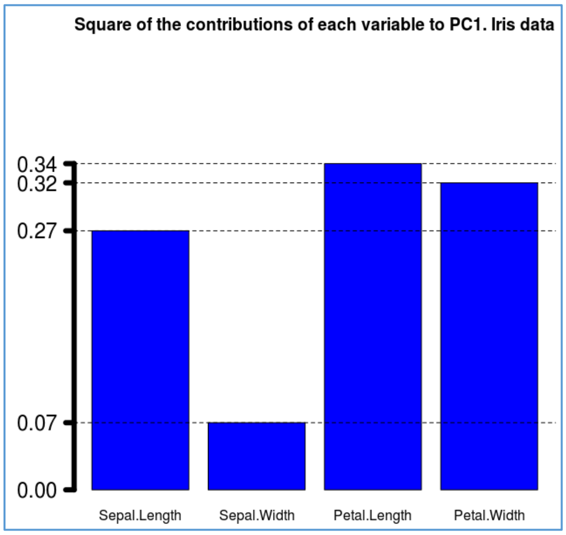

Figure 1. Square of the contribution of each variable to PC_1 in Iris data. Sum of squares equals 1.

Figure 1 shows the square of the contributions of each of the original variables to PC_1. The values on the Y-axis are the squares of the values in Table 1. This results in: 0.27+0.07+0.34+0.32=1.

In PCA, the contributions will be larger for those variables whose directions is more similar to the direction of PC_1. The direction of the largest variance (direction of PC_1) is not equal to the direction of the original variance with largest variance for two reasons: (1) The data was standardized to variance 1 for all variables, and (2) the original variables are to some extend correlated with each other.

Estimation

This section gives an overview of the mathematical operations that enable to estimate the PCs given a centered and standardized data matrix. This section consists of three parts: (A) Errors in the PCA models, (B) Estimation of PCs via eigen-decomposition and (C) Estimation of PCs via singular value decomposition.

A) Errors in the PCA models

Equation 1 showed how to extract PC_1 as a linear combination of the original variables. If we wanted to go back and obtain the original coordinates of our data $\mathbf{X}_{\mathbf{ixj}}$ based on the PC_1 coordinates ($\mathbf{t}_1$) and the loadings $\mathbf{p_1}$, we could do it as follows:

$$\mathbf{X}= \mathbf{t}_1\ \mathbf{p_1}^{\mathbf{T}} + \mathbf{E}_1 \qquad \textit{Equation 5}$$

Where $\mathbf{E_1}_{\ \mathbf{ixj}}$ is matrix containing the residuals that are made when trying to predict the coordinates of the original variables based on the PC_1 coordinates and the coefficients that relate the original variables to PC_1. The reason why $\mathbf{E_1} > 0$ is that PC_1 is a simplification of the original variables. $\mathbf{E_1}$ is the errors that we make when we try to summarize four dimensions with only one. To achieve a fairer representation of the original data, we may need to extract more PCs. Equation 6 shows how PC_2 is extracted after PC_1.

$$\mathbf{t}_2= \mathbf{p}_2\mathbf{E_1} \qquad \textit{Equation 6}$$

Where $\mathbf{t_2}$ is a column vector containing the coordinates of the samples on PC_2 and $\mathbf{p_2}$ is a column vector containing the coefficients that allow to express PC_2 as a linear combination of the original variables. $\mathbf{t_1}$ and $\mathbf{t_2}$ can be combined into a $\mathbf{T}_{\mathbf{1,2}\ \mathbf{ix2}}$ matrix containing the coordinates of the samples in the 2-dimensional space defined by PC_1 and PC_2. For this space the samples will show a large variation since we are using the first two linear combinations as variables. Furthermore the data will show more variation for PC_1 than for PC_2.

In the Iris data we can compare the variation in the data when this is plotted in PC_1 vs PC_2 and when it is plotted in two of the original variables.

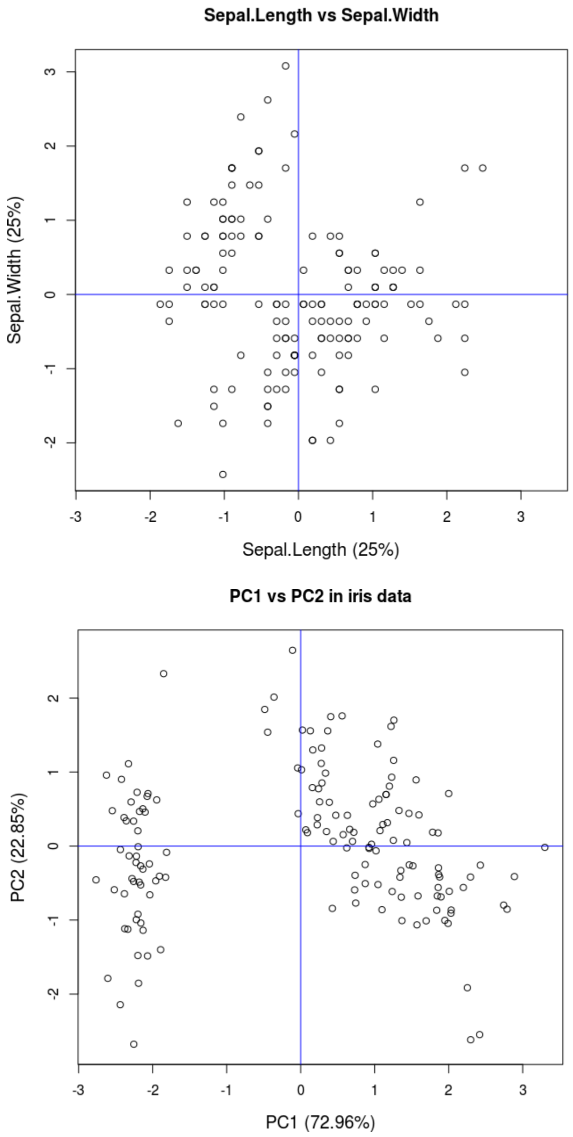

Figure 2. Comparison between plot with original variables and plot with PCs, for Iris data.

In Figure 2 we observe that the data shows more variation in the PC1 vs PC2 plot. Furthermore, in this plot the data is more spreaded across the X-axis (PC1) than across the Y-axis (PC2). In the plot of Sepal.length vs Sepal.width, the spread of the samples is similar across both axes because both of the variables have the same variance due to standardization.

When we want to improve the interpretability, or if we aim to remove some noise, we can select only the first *k* PCs. Then, Equation 2 becomes:

$$\mathbf{T}_{\mathbf{1:k}}=\mathbf{XP} + \mathbf{E}_{\mathbf{k}} \qquad \textit{Equation 7}$$

Where $\mathbf{T}_{\mathbf{1:k}\ \mathbf{ixk}}$ is a matrix with the coordinates of the samples in the space defined by the first *k* PCs, where the samples are in rows and the PCs in columns. $\mathbf{P}_{\mathbf{kxk}}$ is the loadings matrix for the first *k* PCs, and $\mathbf{E}_{\mathbf{k}\ \mathbf{ixk}}$ is the residuals matrix when *j-k* PCs are left out. In other words, $\mathbf{E}_{\mathbf{k}}$ are the errors that we make on the sake of simplicity. Note that if *k=j*, then $\mathbf{E}$=0, and $\mathbf{X}$ could be obtained as in Equation 2.

B) Estimation of PCs via eigen-decomposition

In part (A) we saw how the PCs relate to the original variables and to the previous PCs extracted. However, we have not discussed yet how each of the PCs are extracted. In other words, we have not seen yet how to extract the the linear combinations that fulfil our aim (maximizing the variance) and satisfy the two assumptions. It turns out, that a well-known concept in multivariate statistics called “Eigen-decomposition of a matrix”, allows to achieve this.

The concept of Eigen-decomposition of a matrix is directly related to PCs. However eigen-decomposition is a more general concept. Finding Eigen-decomposition of a square matrix $\mathbf{A}_{axa}$ consists of finding the vectors than are invariant when they are pre-multiplied by the matrix. We call these vectors eigenvectors, and the scalars that express the relationship between these (invariant) vectors before and after the transformation are called eigenvalues. In mathematical notation, the eigen-decomposition of a matrix $\mathbf{A}_{axa}$ can be expressed as:

$$\mathbf{Au}=\lambda\mathbf{u} \qquad \textit{Equation 8}$$

Where $\mathbf{A}_{axa}$ is any square matrix, $\mathbf{u}$ is an eigenvector of $\mathbf{A}_{axa}$ and λ is the corresponding eigenvalue of $\mathbf{u}$. Finding the eigenvectors of a matrix is interesting in matrix algebra because they contain the most relevant information of $\mathbf{A}_{axa}$ in a simplified way (vectors are more easy to interpret that matrices). Thus the set of eigenvectors of a matrix can be regarded as a simplified version of the matrix. For the PCA purpose, the eigenvectors have important properties:

-They point at the direction of largest variance (aim of PCA)

-They are orthogonal to each other (assumption 1)

-Their sum of squares is equal to 1 (assumption 2)

It is common to present the eigen-decomposition of a matrix as a matrix of eigenvectors ($\mathbf{U}_{jxc}$) and a diagonal matrix with the eigenvalues $\mathbf{\Delta}_{cxc}$. The matrix $\mathbf{U}_{jxc}$ should have the eigenvectors as columns, and the eigenvectors should be sorted based on the eigenvalues (the eigenvectors with the largest eigenvalue should be on the first column, and so on). $\mathbf{U}_{jxc}$ will have as many rows as original variables we have (*j*) since they are coefficients. And the number of columns (*C*) corresponds to the number of eigenvectors of $\mathbf{A}_{axa}$ and is equal to its rank. The eigenvalues in the $\mathbf{\Delta}_{cxc}$ matrix should be sorted from largest (upper left) to lowest (bottom right). The eigenvalue of a vectors $\mathbf{u}$ correspond to the variance that the matrix*[1] $\mathbf{A}_{axa}$ have in the direction defined by $\mathbf{u}$. Thus, by sorting both the eigenvectors and their eigenvalues in their corresponding matrices we have sorted the PCs based on their variance.

Considering the full eigen-decomposition, equation 8 can be expressed as follows:

$$\mathbf{AU}=\mathbf{\Delta U}\text{; thus: }\mathbf{A}=\mathbf{U\Delta U}^{-1} \qquad \textit{Equation 9}$$

Since the eigenvectors satisfy the aim and the two assumptions, we can say that the eigenvectors are the PCs. Furthermore, since the PCs are sorted as columns in the $\mathbf{P}$ matrix and the eigenvectors are sorted as columns in the $\mathbf{U}$ matrix, we can say that $\mathbf{U}$ is equal to $\mathbf{P}$. Thus, in order to estimate the PCs, one could think of applying Equation 9 to our data matrix in order to estimate $\mathbf{U}$ and $\mathbf{\Delta}$. However, $\mathbf{A}_{axa}$ cannot be replaced by $\mathbf{X}_{ixj}$ because $\mathbf{A}_{axa}$ needs to be square. A possible solution is to extract the eigenvectors of the cross product $\mathbf{X}^{\mathbf{T}}\mathbf{X}_{jxj}$ instead since this is square*[2]. Thus, replacing $\mathbf{A}_{axa}$ by $\mathbf{X}^{\mathbf{T}}\mathbf{X}_{jxj}$ in equation 9. And advantage of such replacement is that $\mathbf{A}$ becomes symmetric. When $\mathbf{A}$ is symmetric, Equation 9 can be simplified as:

$$\mathbf{A}=\mathbf{U\Delta U'} \qquad \textit{Equation 10}$$

This is a considerable simplification in matrix algebra. Thus, Using the cross product $\mathbf{X}^\mathbf{T}\mathbf{X}$ is advantageous.

It is possible to find the eigen-decomposition of a matrix by trying different combinations of λ and $\mathbf{u}$ in Equation 8. However, this is not practical and the eigen-decomposition of a matrix is often done by deriving the characteristic equation, that is not explained here.

After applying the characteristic equation on the Iris data, we obtain $\mathbf{U}$ and $\mathbf{\Delta}$:

Formula 2. Results of applying the characteristic equation on the Iris data

$$\mathbf{U}=\mathbf{P}=\begin{bmatrix} 0.53 & -0.38 & 0.72 & 0.26 \\ -0.27 & -0.92 & -0.24 & -0.12 \\ 0.59 & 0.02 & -0.14 & 0.8 \\ 0.57 & 0.07 & -0.63 & 0.52 \end{bmatrix} \qquad \mathbf{\Delta}=\begin{bmatrix} 2.92 & 0 & 0 & 0 \\ & 0.91 & 0 & 0 \\ & & 0.15 & 0 \\ Sym & & & 0.02 \end{bmatrix}$$

If we sum the elements in the diagonal of $\mathbf{\Delta}$ we obtain 4, which is same as the variance of the data matrix $\mathbf{X}$ once it was standardized to each variable having variance 1. Thus, the total amount of variance of the data is the same in the 4d-space defined by the original space and in the 4d-space defined by the PCs.

If we divide the first element in the $\mathbf{\Delta}_{\mathbf{4x4}}$ matrix (2.92) by the total variance explained by the PCs (4), we obtain 0.73, which is the portion of variance explained by PC_1.

C) Estimation of PCs via singular value decomposition

The PCs can also be estimated via singular value decomposition (SVD). SVD implies that a given matrix can be discomposed in two matrices of orthogonal vectors and a diagonal matrix with what are known as singular values.

$$\mathbf{A=PSQ'} \quad \textit{Equation 11}$$

Where $\mathbf{P}$ and $\mathbf{Q}$ are two matrices of orthogonal vectors as columns and $\mathbf{S}$ is a diagonal matrix of singular values. Unlike $\mathbf{U}$, $\mathbf{P}$ and $\mathbf{Q}$ are not unique solutions for $\mathbf{A}$. However, $\mathbf{P}$ and $\mathbf{Q}$ are different forms of writing $\mathbf{U}$**.** When we apply SVD to the cross product $\mathbf{X}'\mathbf{X}$, replacing $\mathbf{A}$ by $\mathbf{X}$ in equation 11, we get:

$$\mathbf{X'X=(PSQ')'PSQ'=QSP'PSQ'} \quad \textit{Equation 12}$$

Since $\mathbf{P}$ is orthogonal ($\mathbf{P'P=I}$), we get:

$$\mathbf{Q\,S^2Q'} \quad \textit{Equation 13}$$

When SVD is applied to a symmetric matrix as the cross-product $\mathbf{X}'\mathbf{X}$, $\mathbf{Q}$ is equal to $\mathbf{U}$**.** We observe that both eigen-decomposition (Equation 10) and SVD (Equation 12) yield the same (unique) eigenvectors. Furthermore, the eigenvalues ($\mathbf{\Delta}$) are equal to the square singular values ($\mathbf{S}^2$).

Visualization

In PCA analysis it is common to show the pairwise scatterplots with the coordinates of the samples in the first PCs (scores plots). For the Iris data we have four PCs. If we focus on the first three, we have a total of six scatterplots:

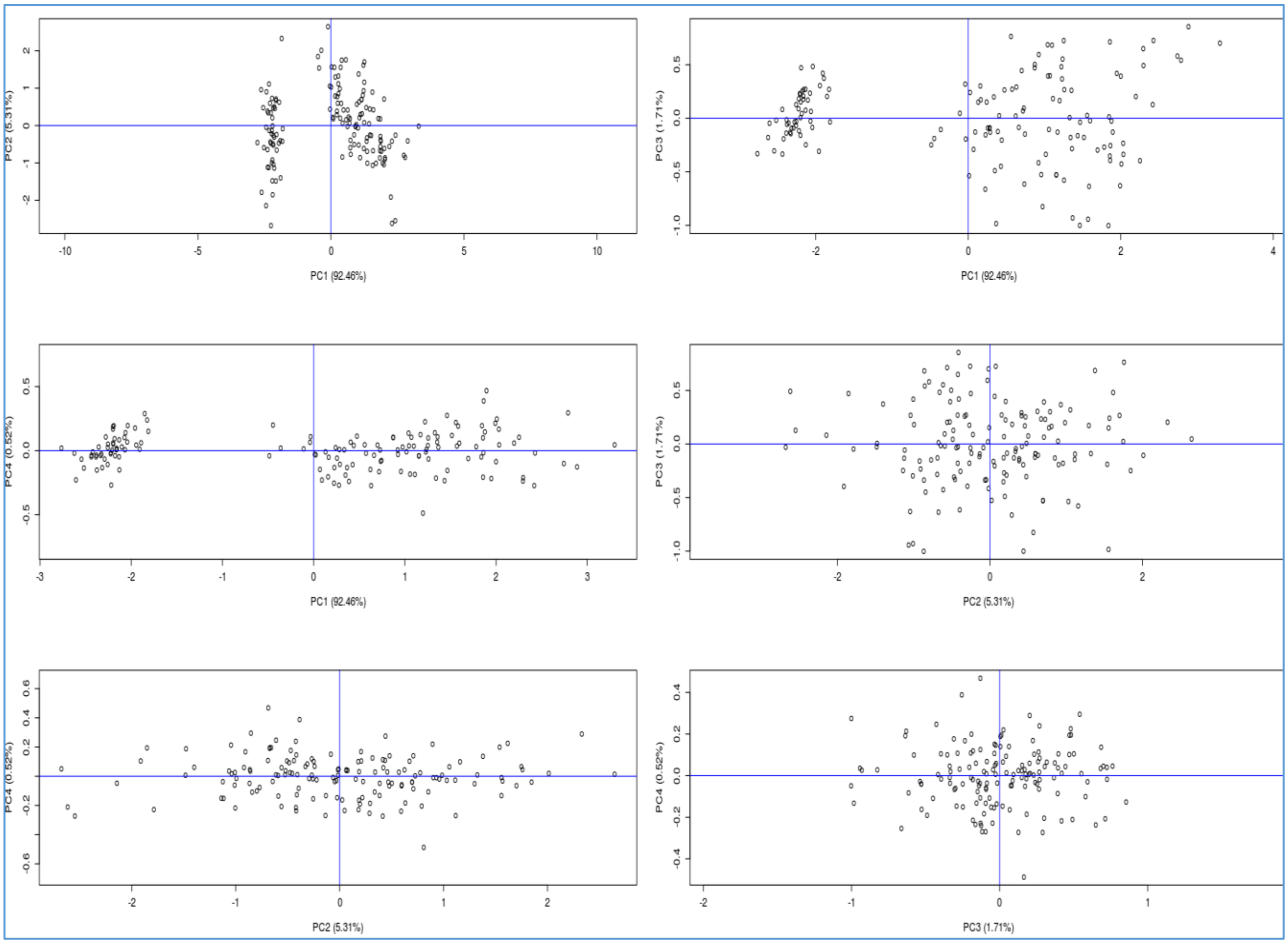

Figure 3. Scatterplots for the first pairs of PCs. Iris data

It is also possible to visualize the data simultaneously in three PCs or even more. For the Iris data, it is interesting to compare the spread of the data when we use the coordinates of three of the original variables and the coordinates of the data on the first PCs, as shown in Figure 4.

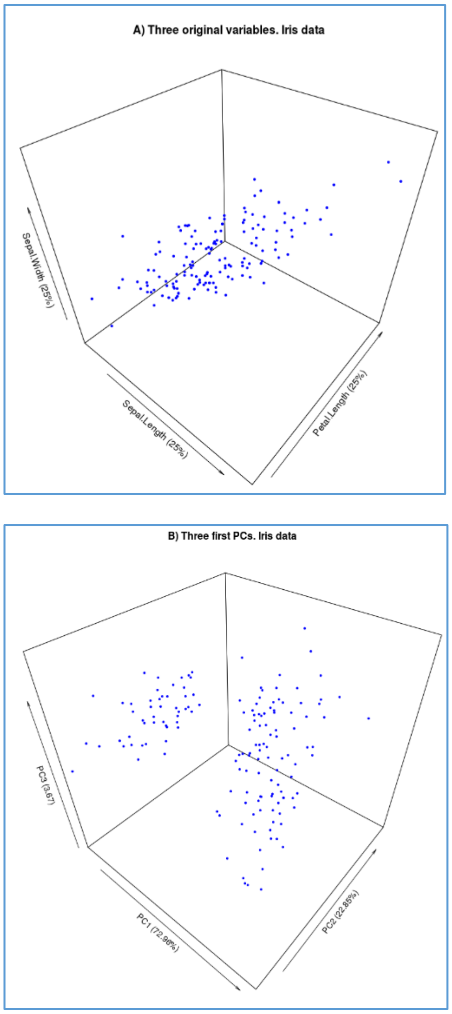

Figure 4.Comparison of variance explained between three of the original variables and the first three PCs.

In Figure 4a we see that most of the variation across samples occurs in a direction that is different to the directions of the original variables. Furthermore, the variation across samples is very similar in the three axes. In Figure 4b, however, PC_1 captures the largest possible variation across samples, distinguishing between two groups of samples. Furthermore, there is much more variation in the direction of PC_2 than for PC_3.

Another type of plots in PCA analyses are the loadings plots that show the contribution of each of the variables for a pair of PCs (generally PC_1 and PC_2). This allows to see how alike the variables are for the pair of PCs selected. Note that these plots consist of plotting the coefficients that relate the original variables to the PCs (rows of the $\mathbf{P}_{\mathbf{jxj}}$ matrix). For the Iris data the loadings plot for PC_1 and PC_2 is shown in Figure 5.

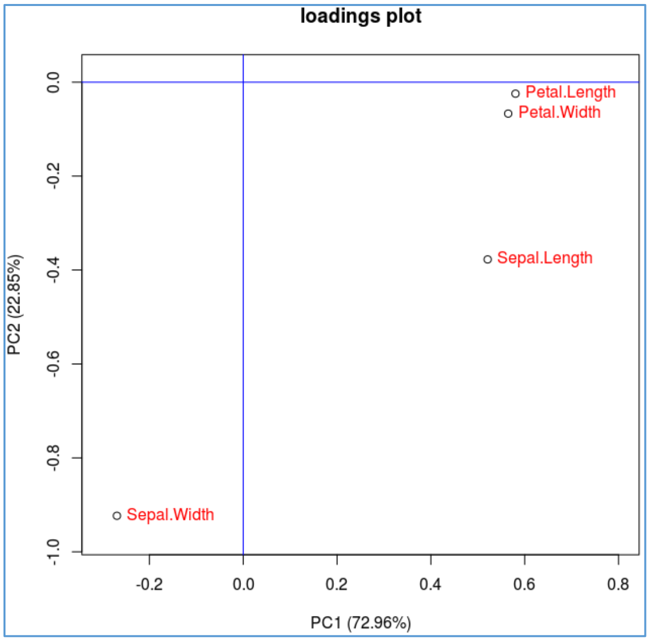

Figure 5. Loadings plots

In Figure 5 we see that two of the variables (Petal.Length and Petal.Width) are very similar with respect to the first two PCs. Furthermore, Sepal.width is very different from the other original variables.

It is also common to combine the scores and loadings plots in what are called “biplots”. For the Iris data the biplot for PC_1 and PC_2 is shown in Figure 6.

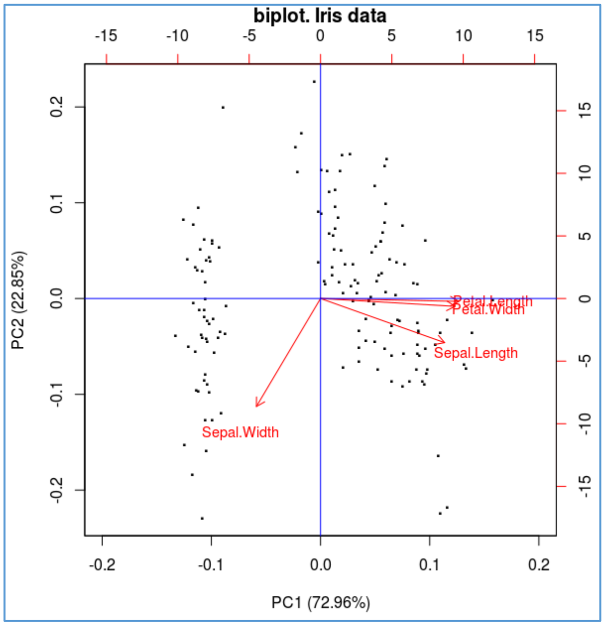

Figure 6. Biplot for iris data

In Figure 6, we observe that the two variables that are more alike (Petal.Length and Petal Width) have a direction that is very similar to the direction of the first PC, which is same as the direction in which two groups of samples are distinguished. PC_2 is in the direction that best distinguished within the two large groups. Among the original variables, Sepal.width is the variable that better achieves the within groups separation. When we look at the coefficients that relates Sepal.width to PC$_2$, we observe that it is very strong (-0.92), as shown in the *[2,2]* element of matrix $\mathbf{U}$ in Formula 2.

Data preprocessing

When estimating the PCs via eigen-decomposition or SVD, the data is usually centered. This is because based on the equations used, the new axis will be given as points and each of these points make an axis when they are connected to the origin. If the data is not centered the axis will be chosen based on where the data falls within the space, rather on where the data shows the maximum variation. Figure 7 shows a biplot for PC$_1$ and PC$_2$ in the Iris data when the data matrix was not centered.

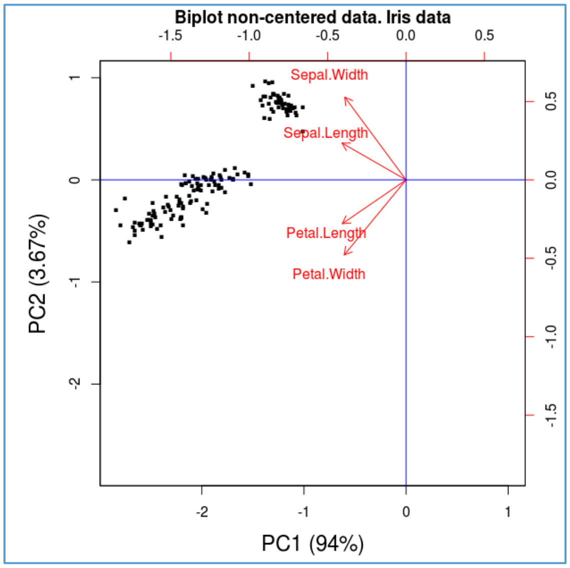

Figure 7. Biplot for noncentered data

Figure 7 shows that PC$_1$ points in a direction that is close to the centroid. The further the centroid is from the origin, the closer PC$_1$ will point towards the centroid.

Another important aspect of data preprocessing for PCA is the standardization of the variables to variance 1. If this preprocessing was not done, the direction of PC1 will be governed by the original variables with largest variance and the other variables that are strongly correlated with it. Figure 8 shows a biplot for the Iris data (PC$_1$ vs PC$_2$) when the data was centered but not standardized:

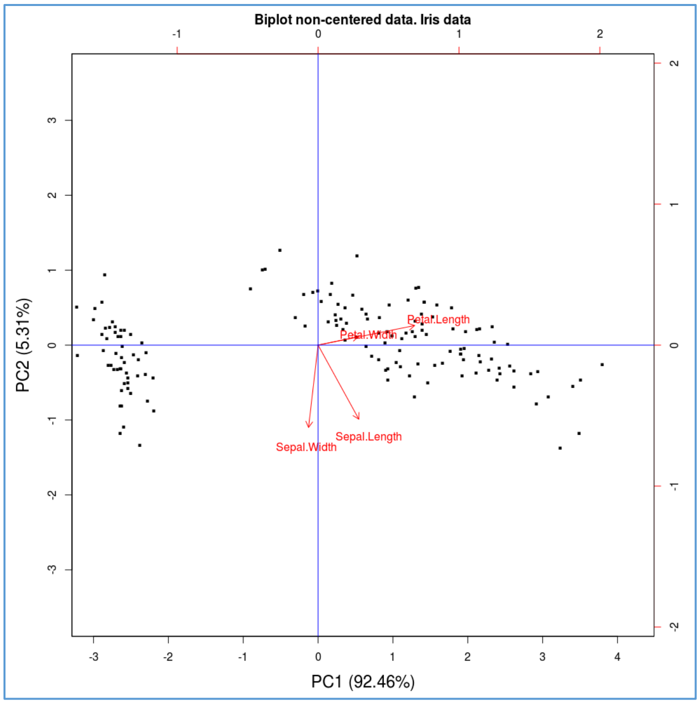

Figure 8. Biplot non-standarized data

In Figure 8 we observe that the direction of the PC₁ is governed by the original variable with largest variance (Petal.Length, as shown in Table 3). This is undesirable, because we should assume that all the original variables are equally important, independently of their variance. Since some variables are more prone to show variation based on the unit measure.

Table 3 provides the variance of the original variables for the Iris data:

Table 3. Variance of the original variables. Iris data

Variable	Variance
Sepal.Length	0.69
Sepal.Width	0.19
Petal.Length	3.12
Petal.Width	0.58

Outliers detection

The so-called "Influence plot" uses the PCA scores and residuals and it is very useful for the identification of outliers. This plot consists of placing statistic Hotelling T^2 2-test statistic along the horizontal axis and the statistics Q Residuals in the vertical axis. Both of these statistics are computed when the first *k* PCs are used. By definition, the Q Residuals statistics will be smaller for large values of *k*, since the more PCs are considered the better we can reproduce the original data.

Hotelling T^2 is an estimate of the score of a sample in the new space. This statistics aims to summarize several coordinates into an scalar. The Hotelling T^2 statistics is computed as the triple product of the score in the space, the variance diagonal matrix of the PCs and the transpose of the score in the space.

$$\mathrm{T_i^2} = \mathbf{t_i^T \Delta t_i} \qquad \textit{Equation 14}$$

Where $\mathrm{T_i^2}$ is an scalar corresponding to the $\mathrm{T^2}$ Hotelling's statistic of sample i. $\mathbf{t_i}$ is a vector of length *k* corresponding to the coordinates of sample *i* in the space defined by *k* PCs. $\mathbf{t_i}$ cooreespnds to the i row of the $\mathbf{T}$ $_{\mathbf{ixk}}$ matrix and $\mathbf{\Delta}$ $_{\mathbf{kxk}}$ is a diagonal matrix with the variance of the PCs.

The Q residuals statistics is a measure of to which extend the original coordinates of a sample can be predicted given the coordinates of the PCs and the loadings matrix ($\mathbf{P}$ $_{\mathbf{jxj}}$).

$$\mathrm{Q_i} = \mathbf{e_i^T e_i} = \mathbf{x_i^T (I - PP^T) x_i} \qquad \textit{Equation 15}$$

Where $\mathrm{Q_i}$ is a scalar corresponding to the Q residual statistic of sample *i*, $\mathbf{e_i}$ is a vector of length *j* with the residuals for sample *i* when we attempt to predict the coordinates of the original variables based on the scores and the loadings matrix. $\mathbf{e_i}$ corresponds to the *i* row of the matrix $\mathbf{E_{k\ ixj}}$ in equation 7. $\mathbf{x_i}$ is a vector corresponding to the *i* row of the $\mathbf{X}$ $_{\mathbf{ixj}}$ data matrix.

Glossary

Variables: variables can be regarded as axes.

PCs: linear combinations of the original variables, new axes, new directions in which the samples are plotted. They conform the "new space" in which the samples are plotted. These axes are "basic axes" in the sense that they are defined as simple as possible for a given direction. They are given as coordinates of points as close as possible to the origin that should be connected to the origin to define the direction).

Scores: Coordinates of the samples in the space defined by the selected PCs

Loadings: matrix containing the weights or coefficients that allow to express the PCs as linear combinations of the original variables. Also called rotation or weights. The matrix of loadings is identical to the matrix of the coordinates of the points that when connected to the origin, define the new axes. The columns of the loadings matrix are the eigenvectors extracted from the cross-product matrix of the data matrix.

Notes

*1

Not clear that it refers to the product of the columns and the rows

*2

This is a very weak argument of why the cross-product is used, but I could not come with a stronger one.

I talk about cross-product instead of variance because I do not understand why the "variance" needs to be maximized, instead of, for instance the "standard deviation".