Principal Component Analysis (PCA)

Introduction

PCA is used to represent a multivariate data set, consisting of n observations and k variables as a low- dimensional plane, usually consisting of two to five dimensions, such that an overview of the data is obtained. This overview may reveal relationships between variables, groups of observations, trends and outliers.
Statistically, PCA finds lines, planes and hyperplanes in $k-$dimensional space that approximate the data as well as possible in the least squares sense. It is easy to see that a line or plane that is the least squares approximation of a set of points makes the variance of the coordinates in the line or plane as large as possible. see Fig 1.

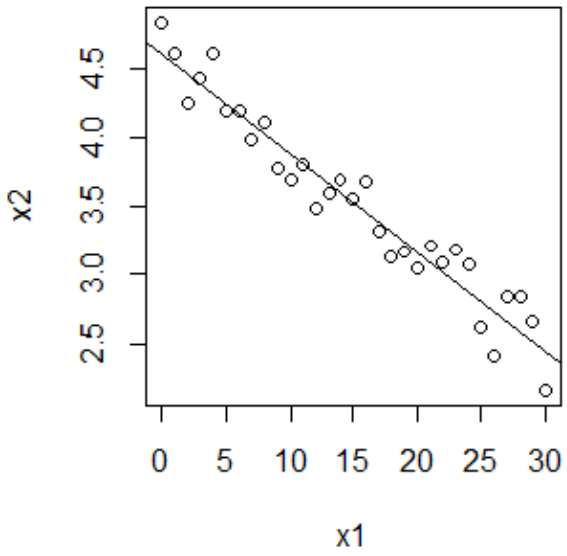

Fig. 1

PCA describes the variation in a set of **correlated** variables $x_1, x_2, ..., x_k$ in terms of a new set of **uncorrelated** variables, $y_1, y_2, ..., y_k$, each of which is a linear combination of the x variables. The new variables are derived in increasing order of importance in the sense that y_1 accounts for as much of the variation in the original data amongst all combinations of $x_1, x_2, ..., x_k$. Then y_2 is chosen to account for as much as possible of the remaining variation,

subject to being **uncorrelated** with y_1 and so on, i.e. forming an **orthogonal** coordinate system. The new variables defined by this process, $y_1, y_2, ..., y_k$ are the **principal components**. When conducting PCA the hope is that the first few components will account for a substantial proportion of the variation in the original variables, $x_1, x_2, ..., x_k$ and can, consequently be used to provide a convenient lower-dimensional summary of these variables. PCA can be used to understand the data set or be used as a data reduction technique whereby the principal components are used as input into some other analysis.

A Geometric Interpretation of PCA

Consider a matrix X with n observations and k variables. For this matrix we construct a variable space with as many dimensions as there are variables. Each variable represents one coordinate axis where each variable has been standardised.
Each observation (each row) of the $X-$matrix is placed in the k dimensional variable space. Consequently, the rows in the data table form a swarm of points in this space (see Fig. 2)

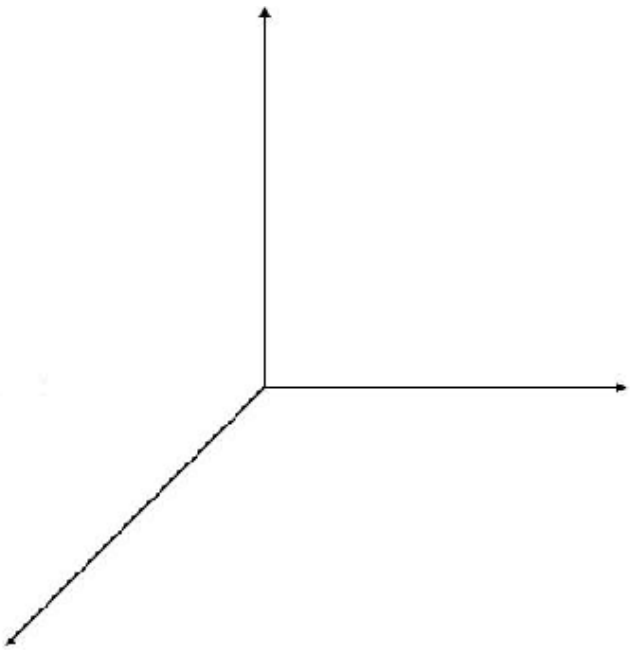

Figure 2 A 3-dimensional variable space, since the variables have been standardised, the vector of variable means is positioned at the origin.

The first principal component (PC1) is the line in k dimensional space that best approximates the data in a least squares sense [N.B. the line that best best approximates the data in a least squares sense is also the line that accounts for maximum variation in the data (see Fig. 1)] This line goes through the mean point at the origin (see Fig. 3). Each observation can now be projected onto this line in order to get a co-ordinate value along the PC1-line. This new coordinate value is known as a **score**.

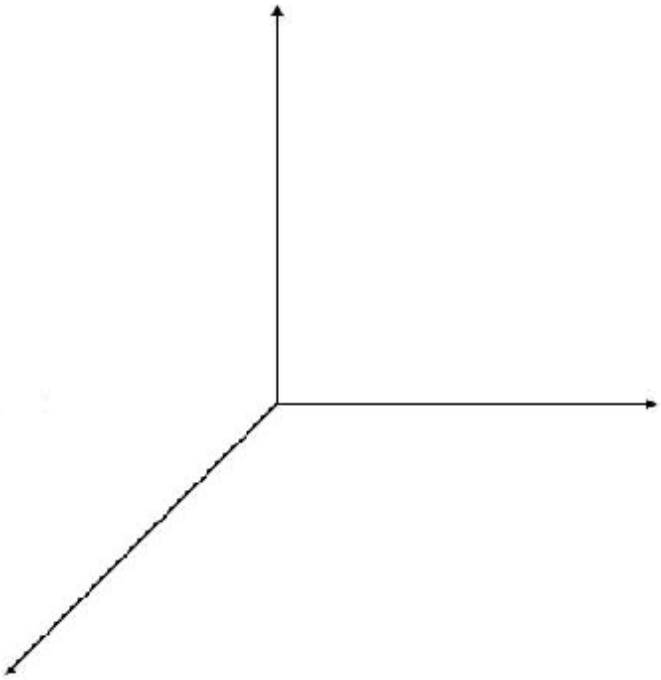

Figure 3. The first principle component, PC1, is the line which best accounts for the shape of the point swarm. It represents the maximum variance direction in the data. Each observation may be projected onto this line in order to get a co-ordinate value along the PC-line. This value is known as a score.

Usually one principal component is insufficient to capture the variation of a data set. Thus, a second principal component, PC2, is calculated. The second PC is also represented by a line in the $k-$dimensional variable space, which is orthogonal to the first principal PC-line (Fig. 4). The line also passes through the mean point at the origin and improves the approximation of the data set as much as possible.

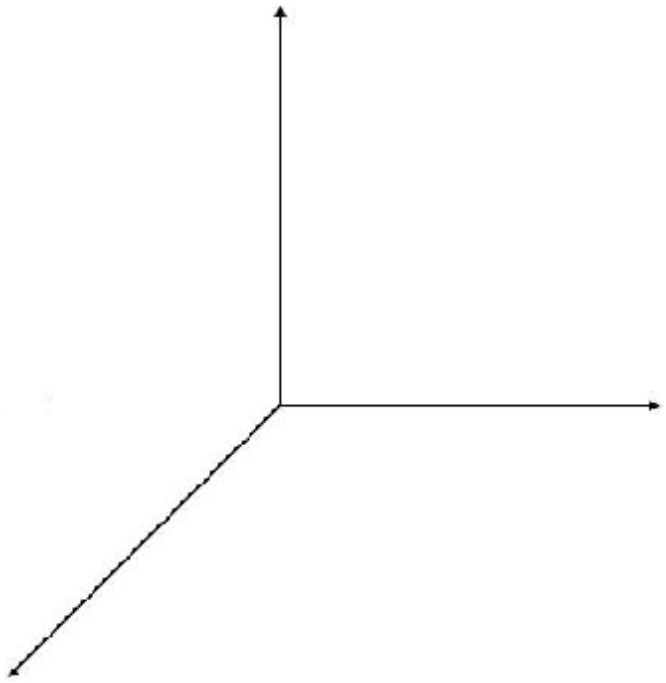

Figure 4 The second principle component, PC2, is orientates so that it captures the second largest source of variation in the data, while being orthogonal to PC1. PC2 also passes through the mean point at the origin.

Two principal components define a model plane

When two principal components have been derived they together define a plane in the $k-$dimensional variable space (see Fig. 5). By projecting all the observations onto this low-dimensional sub-space and plotting the results, it is possible to visualize the structure of the data set. The coordinate values of the observations on this plane are called scores, and hence the plotting of the projected configuration is known as a **score plot**.

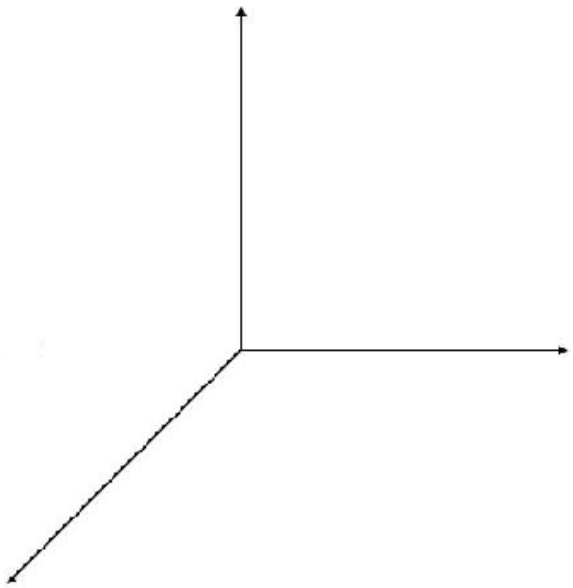

Figure 5 Two PCs form a plane in the k-dimensional variable space. Each observation may be projected onto this giving a score for each.

We will examine the "Foods" data set: the relative consumption of 20 different food items was compiled for 16 countries. The values range from between 0-100% and a high value corresponds to a high consumption. The aim of the study was to investigate the similarities and differences between the 16 countries with respect to their usage of the different food products. PCA creates a condensed summary of the table, which can be analysed graphically by means of the score plot see Fig.6

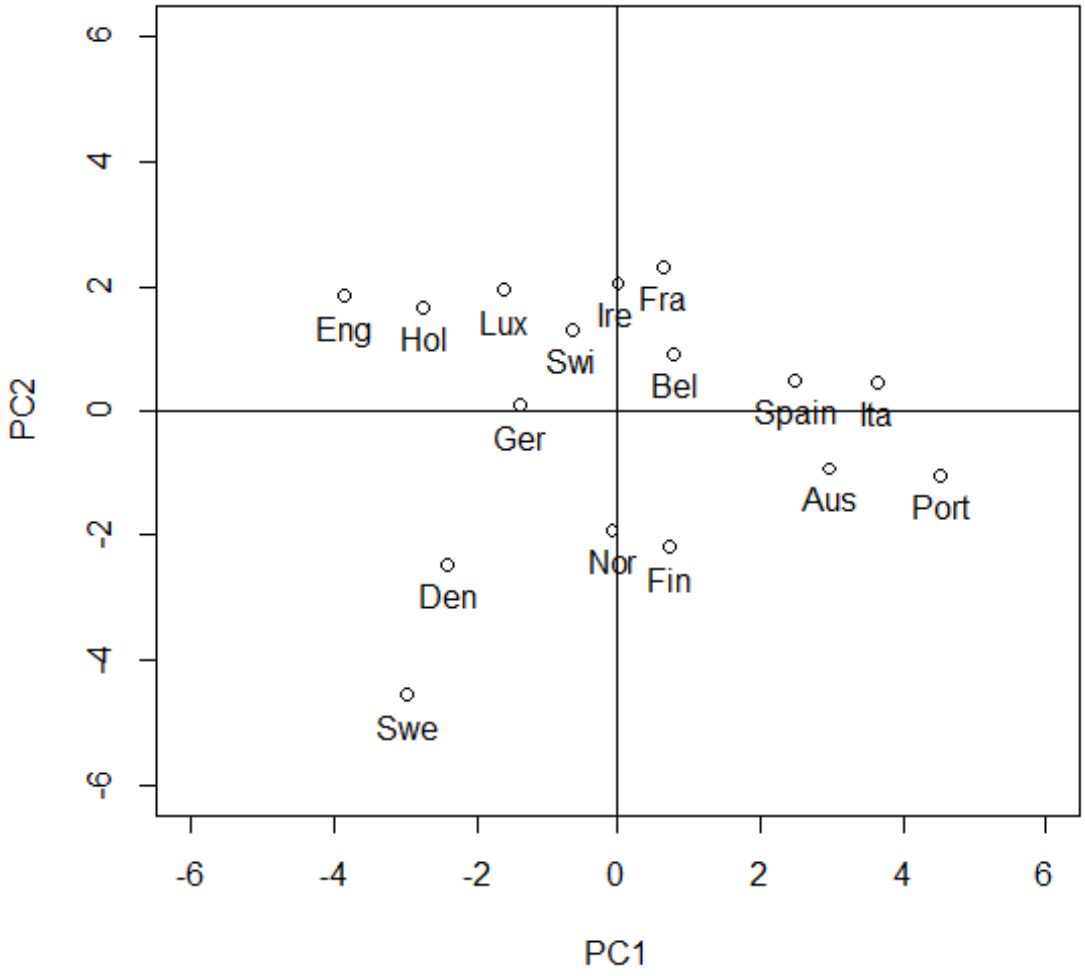

Figure 6. PCA score plot of the first two PCs of the Foods data set. This provides a map of how the countries relate to each other.

Each European country is characterized by two values, one along the first PC and another along the second PC. This score plot is a map of 16 countries. Countries close to each other have similar properties, whereas those from each other are dissimilar with respect to food consumption profiles. The Nordic countries (Finland, Norway, Denmark and Sweden) are located together in the lower left-hand corner, thus representing a group of nations with some similarity in food consumption. Belgium and Germany are close to the origin, which

indicates that they have average food consumption profiles.

How to interpret the score plot

In a PCA model with two components, that is a plane in k-space, we wonder which variables are responsible for the patterns seen among the observations? We would like to know which variables are influential, and also how the variables are correlated. Such knowledge is given by the principal component **loadings** (see Fig. 7)**.**

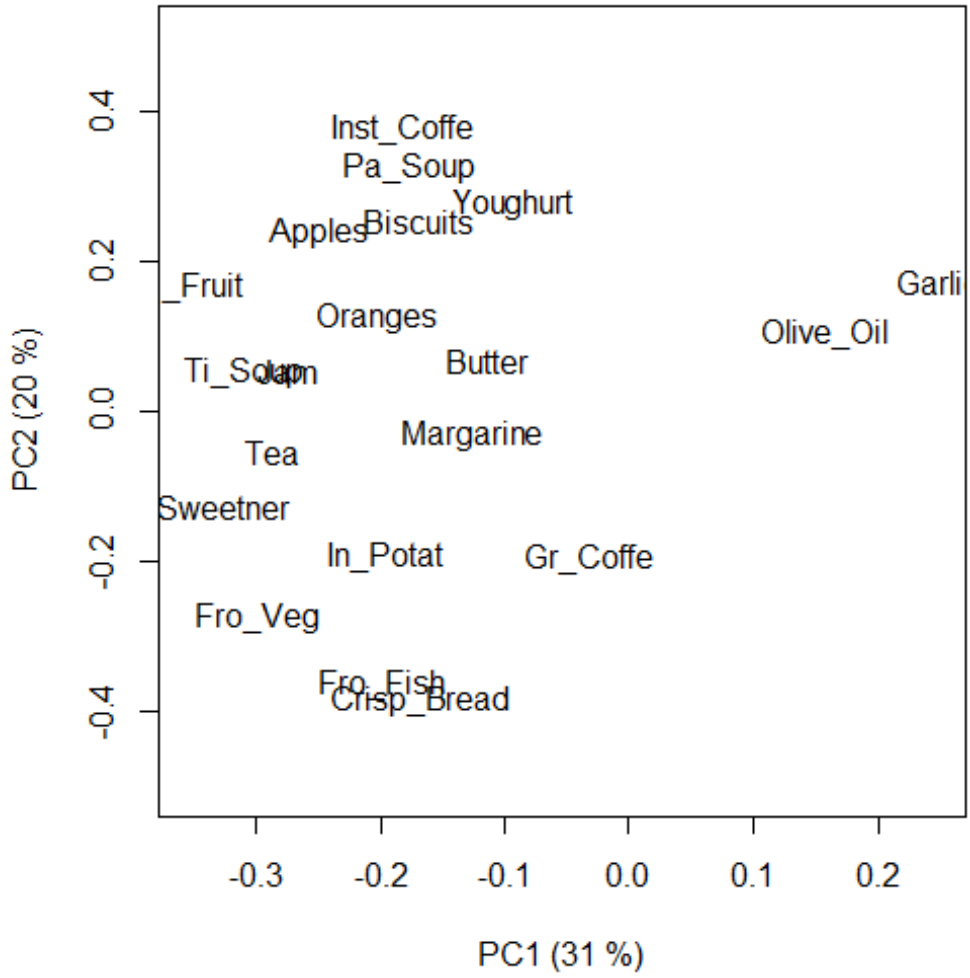

Figure 7. PCA loading plot of the first two principal components of the Foods data set.

The loadings define the orientation of the PC plane with respect to the original X variables. Figure 7 displays the relationships between all 20 variables at the same time. Variables

contributing similar information are grouped together, that is, they are correlated. Crisp bread and frozen fish are examples of two variables that are positively correlated. When the numerical value of one variable increases or decreases, the numerical value of the other variable has a tendency to change in the same way. When variables are negatively correlated they are positioned on opposite sides of the plot origin, in diagonally opposed quadrants. For instance, the variables garlic and sweetener are negatively correlated, meaning that when garlic increases, sweetener decreases and vice versa. Furthermore, the distance to the origin also conveys information. The further away from the plot origin a variable lies, the stronger the impact that it has had on the model. This means for instance, that the variables crisp bread, frozen fish, frozen vegetables and garlic separate the four Nordic countries from the others. The model interpretation suggests that countries like Italy, Portugal, Spain and Austria have high consumptions of garlic and low consumption of sweetener, tinned soup and tinned fruit.

The scores plot and the loadings plot can be overlaid to give a biplot (see Fig. 8)

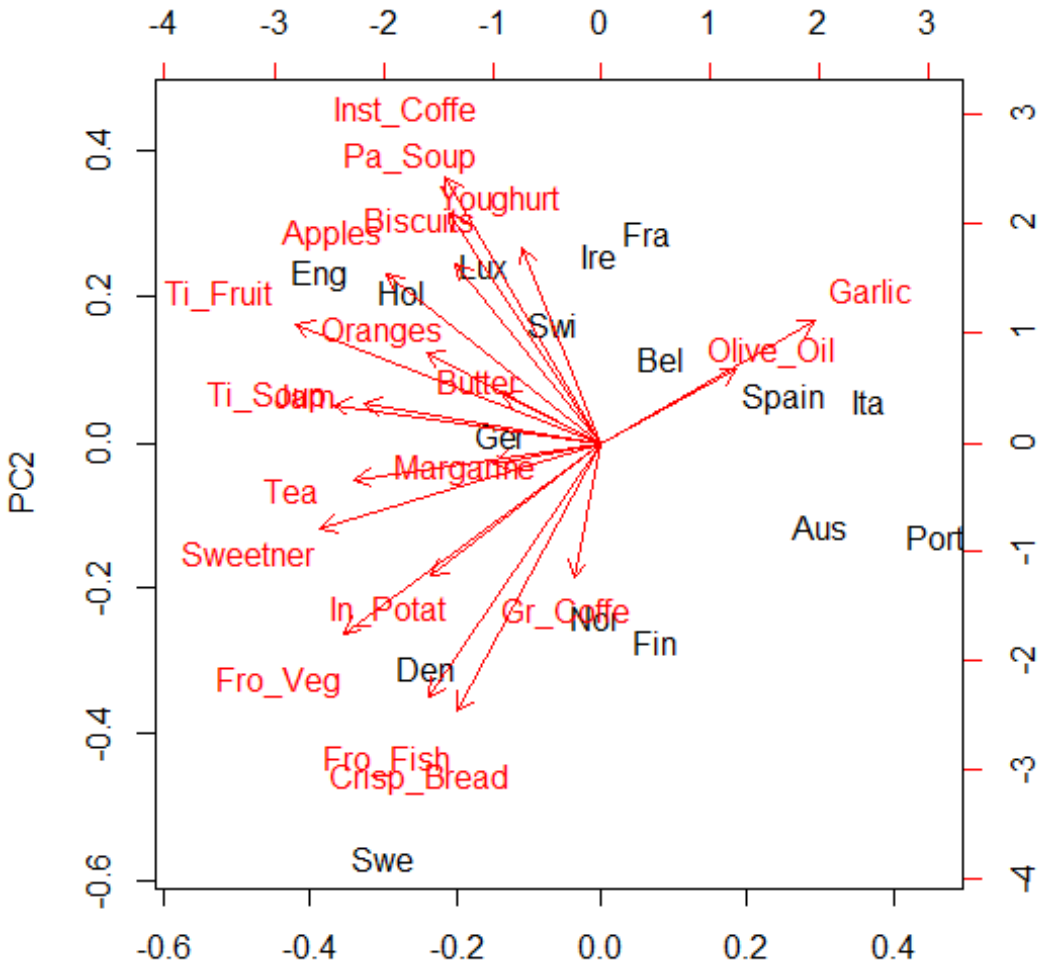
-4
-3
-2
-1
0
1
2
3
Inst_Coffe
Pa_Soup
Youghurt
Biscuits
Apples
Fra
Ire
Lux
Eng
Ti_Fruit
Hol
Garlic
Oranges
Swi
Bel
Olive_Oil
Butter
Spain
Ita
Ger
Margarine
Tea
Aus
Port
Sweetner
In_Potat
Gr_Coffe
Nor
Fin
Fro_Veg
Den
Fro_Fish
Crisp_Bread
Swe
PC2
0.4
0.2
0.0
-0.2
-0.4
-0.6
3
2
1
0
-1
-2
-3
-4
-0.6
-0.4
-0.2
0.0
0.2
0.4

Figure 8