Principal Component Analysis III

Model Diagnostics

How many principal components do we need to describe our data?

Frequently, one or two principal components are not enough to adequately capture the information in the data set. In such cases the descriptive ability of the PCA model improves by using more principal components. There are several approaches that can be used to evaluate how many principal components are necessary, we consider four of these approaches below:

- The Kaiser Criterion
- The Scree Plot
- The proportion of variation accounted for by each principal component
- The proportion of variation accounted for by a PCA model with m components (a measure of the *goodness of fit*, often denoted as $R^2_m X$)

The Kaiser Criterion The Kaiser criterion suggests that we retain all components with eigenvalues greater than 1. Recall that since the data used in PCA is standardised, each variable contributes one unit of variance to the overall variance of the data set. Any component with an eigenvalue greater than one is accounting for a greater amount of variation than one variable and should be retained. Any component with an eigenvalue of less than one is accounting for less variance than has been contributed by one variable, thus should not be retained. This method is simple to apply, however it does not always result in the correct number of components being retained, for example, if one component has an eigenvalue of 1.01 and another has a value of 0.99, should you retain one and not the other?

The Scree Plot The scree plot is a bar or line plot of the eigenvalues associated with each principal component. The idea is to retain components that are contributing significantly to the variance. We look for a large drop in the eigenvalues.(see Fig1).

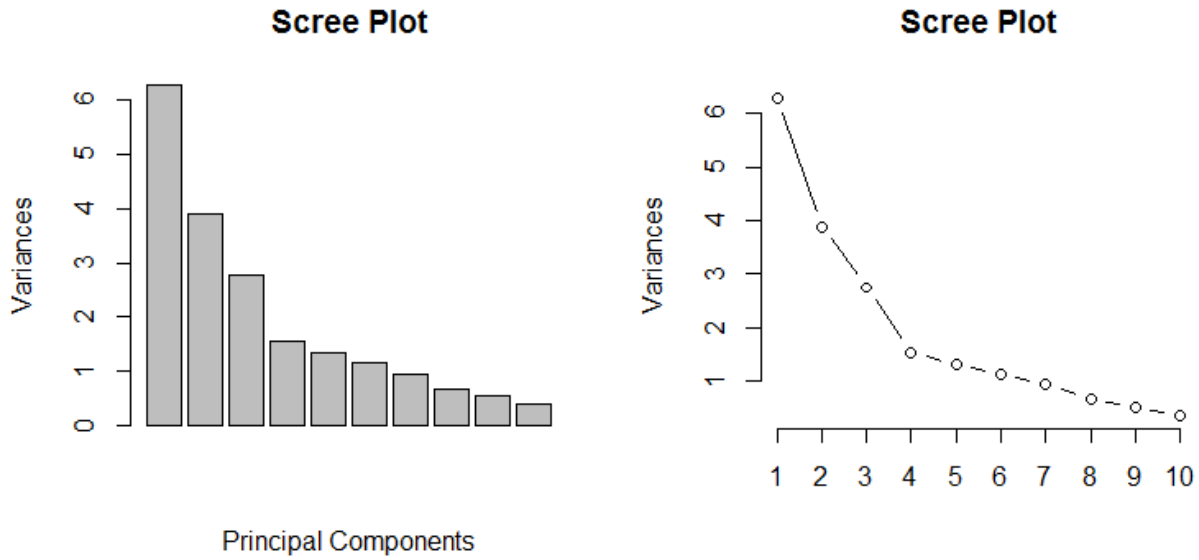

Figure 1. Scree plot for the PCA for the Foods data set

The scree plot is subjective, the plot shown in Figure 1 may indicate that we retain 3 PCs.

The Proportion of Variation Accounted For (R_m^2X) We could also select PCs by setting a minimum threshold for the amount of variance a component accounts for. For example, only retain components that account for 5% or more of the variation. Similarly we could set a threshold for the total amount of variance accounted for by the PCA model. For example, we may require that the PCA model accounts for at least 95% of the variance, in which case we would look at the proportion of the variation explained by a PCA model with m components, denoted by R_m^2X.

R_m^2X represents the proportion of variation accounted for by a PCA model with m components.

In the example with just two variables , x_1 and x_2 we found that with just one principal component $R_1^2X = 0.96$ and with two components, clearly $R_2^2X = 1$.
Note that we calculated R_1^2X and R_2^2X, by considering the ratio of the cumulative sum of the eigenvalues associated with each principal component ($\frac{\lambda_1}{trace(S)}$). Another way of calculating

the proportion of variance explained by a PCA model with m components would be to consider the residuals. Using PCA with m components, a scaled data set $\mathbf{X}$ is modelled as:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

where $\mathbf{T}$ represents the matrix of scores, $\mathbf{P}$ represents the matrix of loadings (or rotations) and $\mathbf{E}$ represents the residuals (or errors).
Remember, the scores represent the position of the observations in relation to the new principal components and the loadings represent the eigenvectors used to define the new PCs. Note that the eigenvectors give the correlation between the original variables and the PCs.

The residuals, $\mathbf{E} = \mathbf{X} - \mathbf{TP}^T$.
The proportion of variation explained by a PCA model with m components ($R^2_m X$) can be written in terms of the residuals as:

$$\begin{aligned} R^2_m X, &= 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} \\ &= 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^2} \end{aligned}$$

For the data set shown below, we will calculate the proportion of the variation that is explained:

(a) by the full PCA model with two components, $R^2_2 X$

(b) by the PCA model with just 1 component. $R^2_1 X$

We showed that the simple data set

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

which when standardised was

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

Let $\mathbf{X}$ represent this scaled dataset. We can represent this scaled data set. in terms of the two principal components:

$$PC1 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \text{ and } PC2 = \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix}.$$

(a) For the case where we have two principal components, the loading matrix $\mathbf{P} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$.

The scores are simply the projection of the original standardised observations onto the new components, so for a PCA model with two components, the scores are:

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

If we use both components in our model, then

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

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

$$\begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} + 0$$

As expected, $E = 0$, since all the variation in the data is explained by two principal components. Clearly the proportion of variation explained by this PCA model will be 1 since all the variation is accounted for. Check:

$$R^2X = 1 - \frac{\sum_{i=1}^{10}\sum_{j=1}^{10} e_{ij}^2}{\sum_{i=1}^{10}\sum_{j=1}^{10} x_{ij}^2} = 1 - \frac{0}{18} = 1$$

(b) For the case where we use just one PC in our model then

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

where $\mathbf{P} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$, (just PC1) , thus

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

So

$$\mathbf{X} = \mathbf{TP}^{-1} + \mathbf{E}$$

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

$$\begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} = \begin{bmatrix} 0.60 & 0.60 \\ -1.56 & -1.56 \\ 0.67 & 0.67 \\ 0.22 & 0.22 \\ 1.36 & 1.36 \\ 1.04 & 1.04 \\ 0.12 & 0.12 \\ -0.87 & -0.87 \\ -0.35 & -0.35 \\ -1.22 & -1.22 \end{bmatrix} + \mathbf{E}$$

Thus the residuals E are

$$\begin{aligned}
\mathbf{E} &= \mathbf{X} - \mathbf{T}\mathbf{P}^{-1} \\
\mathbf{E} &= \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} - \begin{bmatrix} 0.60 & 0.60 \\ -1.56 & -1.56 \\ 0.67 & 0.67 \\ 0.22 & 0.22 \\ 1.36 & 1.36 \\ 1.04 & 1.04 \\ 0.12 & 0.12 \\ -0.87 & -0.87 \\ -0.35 & -0.35 \\ -1.22 & -1.22 \end{bmatrix} \\
\mathbf{E} &= \begin{bmatrix} 0.1 & -0.1 \\ -0.12 & 0.12 \\ -0.37 & 0.37 \\ -0.01 & 0.01 \\ 0.12 & -0.12 \\ -0.13 & 0.13 \\ 0.42 & -0.42 \\ -0.06 & 0.06 \\ -0.05 & 0.05 \\ 0.11 & -0.11 \end{bmatrix}
\end{aligned}$$

To calculate the proportion of the variation explained by the PCA model we need to calculate the residual sum of squares, $\sum_{i=1}^{n} e_{ij}^2$

$$\sum_{i=1}^{10}\sum_{j=1}^{10} e_{ij}^2 = 0.76$$

$$\begin{aligned} R_1^2 &= 1 - \frac{\sum_{i=1}^{10}\sum_{j=1}^{10} e_{ij}^2}{\sum_{i=1}^{10}\sum_{j=1}^{10} x_{ij}^2} \\ &= 1 - \frac{0.76}{18} \\ &= 1 - 0.04 = 0.96 \end{aligned}$$

So using just one principal component in our model, captures 96% of the variation in the data.

Figure 2 shows the goodness of fit measure $R_m^2 X$ for PCA models of the Foods data set are shown for increasing values of m.

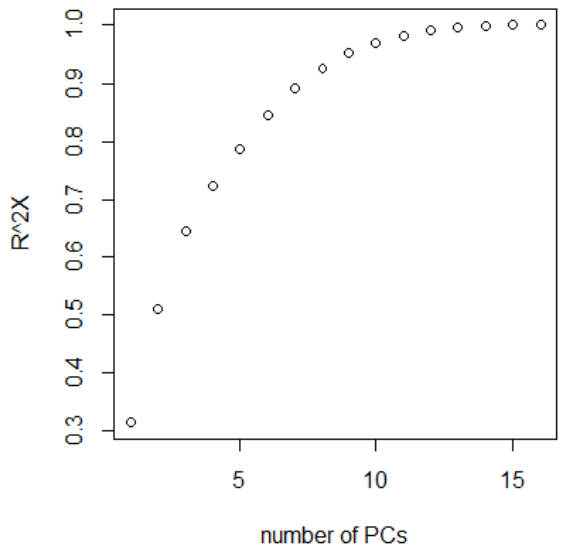

Figure 2.

Outliers

It is possible to detect multivariate outliers in a data set by plotting a confidence ellipse onto a scores plot from a PCA.

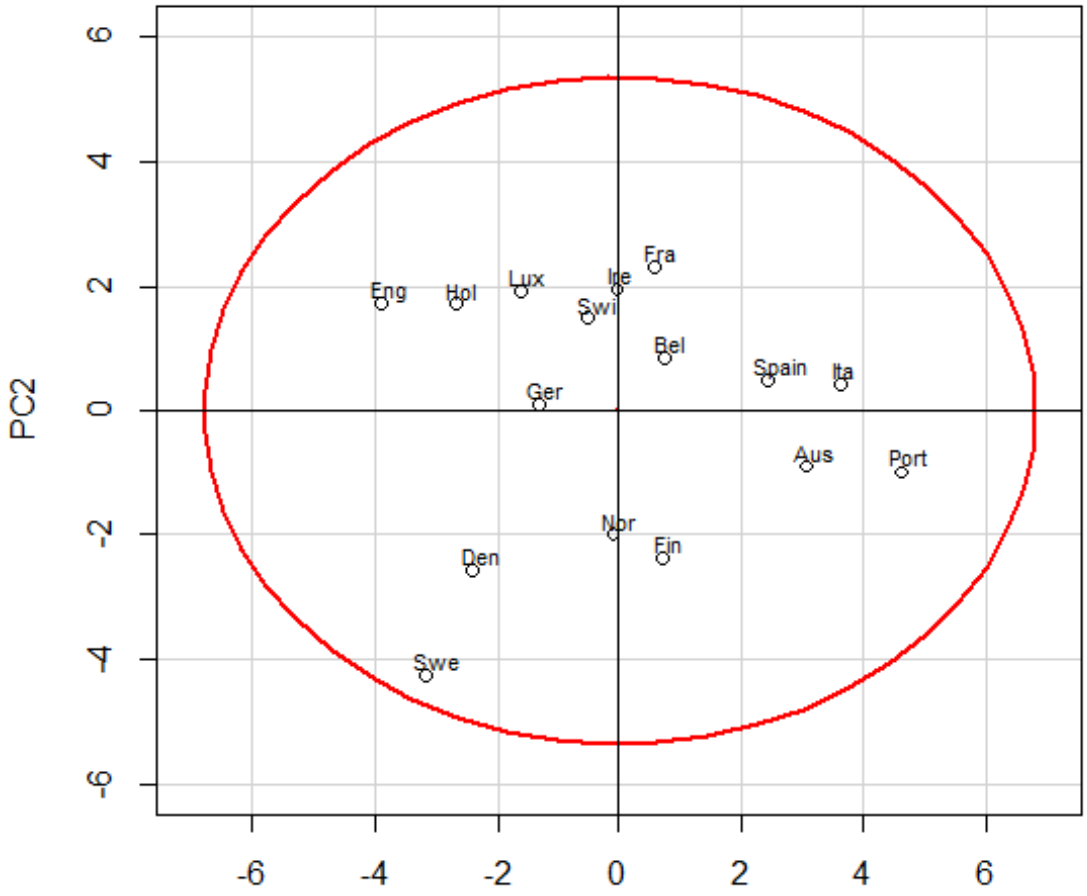

Figure 3 Score plot for the Foods data set with a 95% confidence ellipse.

Figure 3 indicates that there are no outliers in relation to PC1 and PC2.

Which variables are well explained by the PCA model?

To find out the extent to which each variable is accounted for by a PCA model with m components we can examine the residuals. Using PCA with m components, a data set $\mathbf{X}$ is modelled as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

We denote the proportion of the variation of variable j explained by the PCA model (with m components) by $R^2_m X_j$, then

$$\begin{aligned} R^2_m X_j, &= 1 - \frac{\text{residual sum of squares for variable } j}{\text{total sum of squares for variable } j} \\ &= 1 - \frac{\sum_{i=1}^{n} e_{ij}^2}{\sum_{i=1}^{n} x_{ij}^2} \end{aligned}$$

Example

For the same example data set $\mathbf{X}$ shown below, we will calculate the proportion of the variation of variables x_1 and x_2, that is explained:

(a) by the full PCA model with two components $R^2_2 X_1$

(b) by the PCA model with just 1 component. $R^2_1 X_1$

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

(a) For the PCA model with two principal components, the loading matrix $\mathbf{P} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$.

$$\text{and } \mathbf{T} = \begin{bmatrix} 0.85 & -0.14 \\ -2.21 & 0.16 \\ 0.94 & 0.52 \\ 0.31 & 0.02 \\ -1.92 & -0.16 \\ 1.47 & 0.18 \\ 0.16 & 0.59 \\ -1.22 & 0.09 \\ -0.50 & 0.08 \\ -1.73 & -0.16 \end{bmatrix}$$

If we use both components in our model, then

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

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

$$\begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} + 0$$

As expected, $\mathbf{E} = 0$, since all the variation in the data is explained by two principal components. Clearly the proportion of variation explained by this PCA model for variables x_1 and

x_2 will be 1 since all the variation is accounted for. Check:

$$R_2^2 X_1 = 1 - \frac{\sum_{i=1}^{10} e_{i1}^2}{\sum_{i=1}^{10} x_{i1}^2} = 1 - \frac{0}{9} = 1$$

$$R_2^2 X_2 = 1 - \frac{\sum_{i=1}^{10} e_{i2}^2}{\sum_{i=1}^{10} x_{i2}^2} = 1 - \frac{0}{9} = 1$$

(b) For the case where we use just one PC in our model then

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

where $\mathbf{P} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$, (just PC1) , and $\mathbf{T} = \begin{bmatrix} 0.85 \\ -2.21 \\ 0.94 \\ 0.31 \\ -1.92 \\ 1.47 \\ 0.16 \\ -1.22 \\ -0.50 \\ -1.73 \end{bmatrix}$.

So

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

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

$$\begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} = \begin{bmatrix} 0.60 & 0.60 \\ -1.56 & -1.56 \\ 0.67 & 0.67 \\ 0.22 & 0.22 \\ 1.36 & 1.36 \\ 1.04 & 1.04 \\ 0.12 & 0.12 \\ -0.87 & -0.87 \\ -0.35 & -0.35 \\ -1.22 & -1.22 \end{bmatrix} + \mathbf{E}$$

Thus the residuals $\mathbf{E}$ are

$$\begin{aligned}
\mathbf{E} &= \mathbf{X} - \mathbf{T}\mathbf{P}^T \\
\mathbf{E} &= \begin{bmatrix} 0.7 & 0.51 \\ -1.68 & -1.44 \\ 0.3 & 1.03 \\ 0.21 & 0.23 \\ 1.48 & 1.24 \\ 0.91 & 1.17 \\ 0.54 & -0.30 \\ -0.93 & -0.80 \\ -0.41 & -0.30 \\ -1.11 & -1.33 \end{bmatrix} - \begin{bmatrix} 0.60 & 0.60 \\ -1.56 & -1.56 \\ 0.67 & 0.67 \\ 0.22 & 0.22 \\ 1.36 & 1.36 \\ 1.04 & 1.04 \\ 0.12 & 0.12 \\ -0.87 & -0.87 \\ -0.35 & -0.35 \\ -1.22 & -1.22 \end{bmatrix} \\
\mathbf{E} &= \begin{bmatrix} 0.1 & -0.1 \\ -0.12 & 0.12 \\ -0.37 & 0.37 \\ -0.01 & 0.01 \\ 0.12 & -0.12 \\ -0.13 & 0.13 \\ 0.42 & -0.42 \\ -0.06 & 0.06 \\ -0.05 & 0.05 \\ 0.11 & -0.11 \end{bmatrix}
\end{aligned}$$

To calculate the proportion of the variation of each variable explained by the PCA model we need to calculate the residual sum of squares,

$$\begin{aligned}
\sum_{i=1}^{10} e_{i1}^2 &= 0.38 \\
\sum_{i=1}^{10} e_{i2}^2 &= 0.38
\end{aligned}$$

$$R_1^2X_1 = 1 - \frac{\sum_{i=1}^{10} e_{i1}^2}{\sum_{i=1}^{10} x_{i1}^2} = 1 - \frac{0.38}{9} = 1 - 0.04 = 0.96$$

$$R_2^2X_2 = 1 - \frac{\sum_{i=1}^{10} e_{i2}^2}{\sum_{i=1}^{10} x_{i2}^2} = 1 - \frac{0.38}{9} = 1 - 0.04 = 0.96$$

So using just one principal component in our model, captures 96% of the variation of the variable x_1 and 96% of the variation of the variable x_2.

Example

Below are barcharts showing the fraction of the explained variation of the variables of the Foods data set for a PCA model with 4(a) 1 principal component, 4(b) 2 principal components, 4(c) 3 principal components.

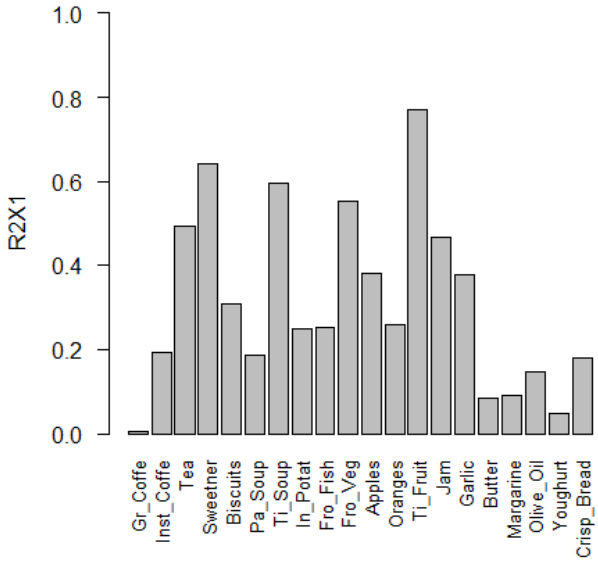

Figure 4 (a) The proportion of variation explained for each variable by a PCA model with 1 component.

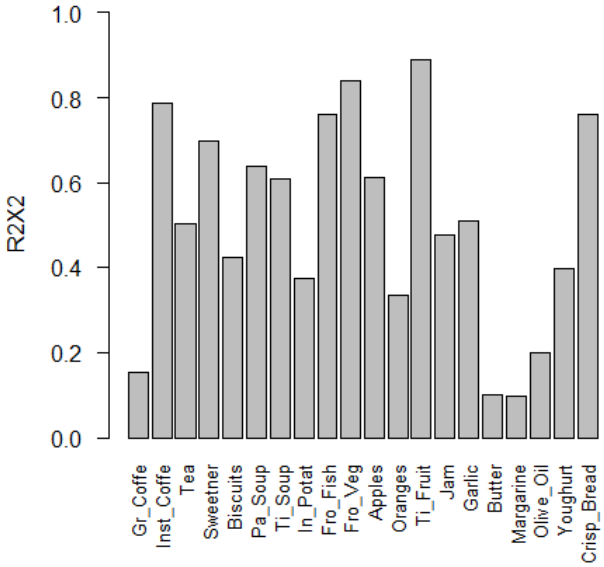

Figure 4(b) The proportion of variation explained for each variable by a PCA model with 2 components.

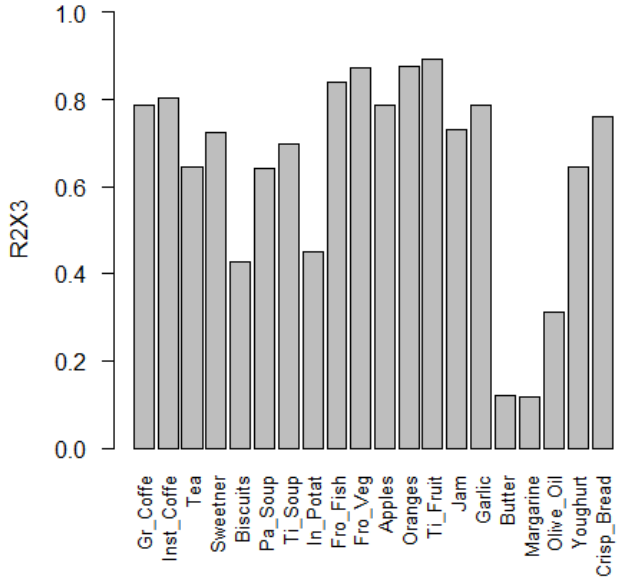

Figure 4 (c) The proportion of variation explained for each variable by a PCA model with 3 components.

Figures 4(a)-4(c) show how the individual variables are modelled by the different principal components. The variable Tin_Fruit is an example of a variable that is well explained by the first PC, and Gr_Coffe is captured by the third PC. There are variables that load onto more than one component, for instance Yogurt which is well explained by the second and third principal components jointly.

Cross Validation

In PCA, the aim of cross validation is to estimate how well the model will generalise to another data set. Just because a model fits the data well, does not mean it is a good model. The basic idea of cross validation is to keep a portion of the data out of the model fitting process (often referred to as a test set) then measure the error when the model is used to predict values from the test set.