Variable removal: Selecting differentiating variables

The dataset originates from the Brazilian Table of Food Composition and has 15 variables and 71 samples of two sources: bovine and chicken. The variables represent water and ash percentage, protein, lipids, cholesterol, chemical elements (Ca, Mg, Mn, P, Fe, Na, K, Cu and Zn), and niacin concentrations.

Tasks:

1. Take a look at the table. What variables would you include and why? Import the variables which look of interest into MATLAB.
2. Does the dataset have missing values? Remove the variable with missing values and remove the observations that have missing values for only one variable, as the dataset is extensive enough to include enough observations if they are deleted.
3. Center and scale the data. Run a PCA model. Plot the explained variance of the principal components. How many principal components will you keep including enough variation for the model?

The dataset contains missing values where the 'Mn' feature was removed, and all other variables were included. In addition, row 8 was deleted because it had a missing value too.

Z-Score was used for centering and scaling of the dataset.

The number of PCs that we chose to include to our model is 8 because it contains more than 95% of the total variation [96,8%].

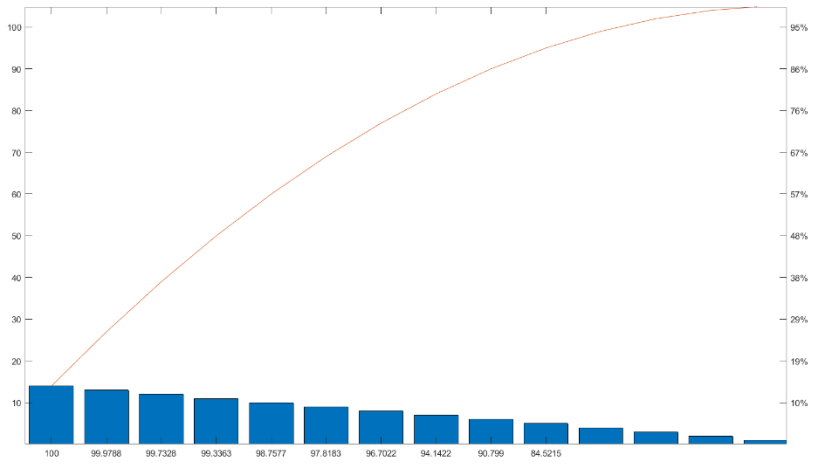

Figure 1. Explained variation by Principal Components.

4. Plot the biplots of the principal components up to the maximum number of principal components you have chosen. Use different colors for the different meat types (beef or chicken). The model tries to capture the whole variation of the dataset. Is there any principal component that distinguishes between the two samples?

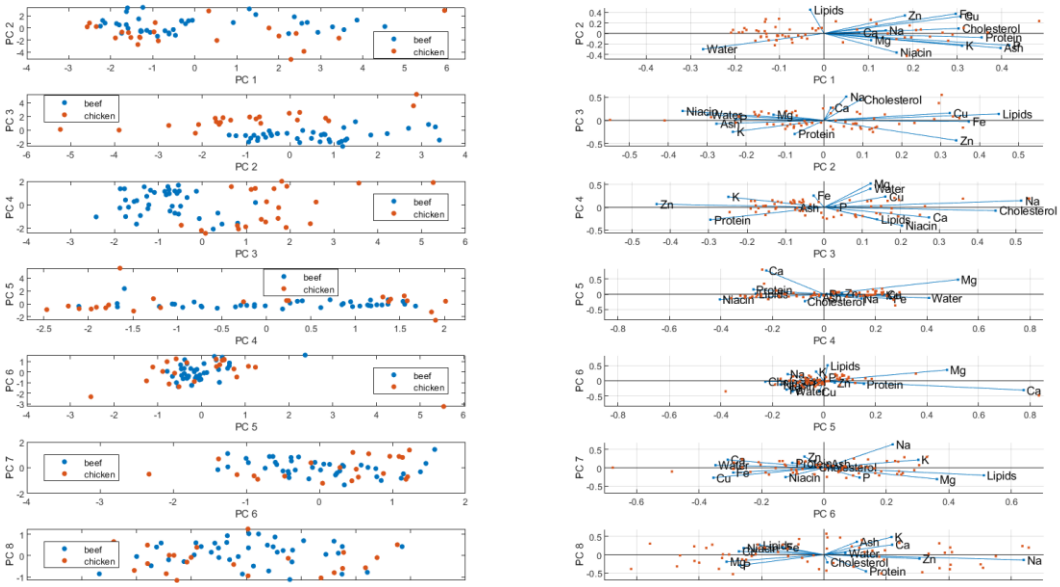

Figure 2. First 7 biplots, with different colors for "beef" and "chicken" types.

5. If you found a principal component able to discriminate analyze the loadings of that PC and the individual scores of it. Which of the “beef” or “chicken” observations are farther away from the PC (have higher scores)?

Figure 3.(a) Loadings of the n-th principal component as a boxplot (b) Scores of the n-th principal component as a function of the observation number, with different colors for "beef" and "chicken".

6. Compute individual PCA models for “beef” and “chicken”. We will call the initial model ***“model 1”***, the chicken model ***“model 2”*** and the beef model ***“model* 3”** for ease. Remember to center and scale the data again using just the samples of each model. Store the mean and standard deviation of the scaling for each model, as you will need it later. Plot the biplots (PC1. Vs PC2) for model 1, model 2 and model 3. Do you see any differences between the direction and magnitude of variable loadings in the components having the most variation? What can you say about models 2-3 compared (a) to each other and (b) to the initial one.

Figure 4 Biplots of PC1 vs. PC2 for (a). Model 1 (b) Model 2 (c) Model 3.

7. Compare the loadings of the first principal component for Model 1, Model 2 and Model 3. Are there any notable differences in them? What are the similarities and dissimilarities between them?

Figure 5. Loadings of the first PC for Model 1, Model 2 and Model 3

8. When doing the individual models for chicken and beef (***Model 2*** and ***Model 3***), you saved the mean and standard deviation of the data matrix for each. Cross-scale the original Model 1 and Model 2 data with the mean and standard deviation from the other model. *E.g. normalize(*XChicken*, 'Center',* meanBeef*, 'Scale',* sigmaBeef*)*.
9. Project the newly scaled "chicken" samples using the ***Model 3*** loadings, resulting in new **T** scores. Plot the biplots for **Model 2** and **3** for all data (both the newly calculated scores and the original model scores, each in a new color). What can you observe about the (a) differences between the new models (b) blend-in of the chicken samples in the beef model and vice-versa.

Figure 6. Biplots of PC1 vs. PC2 for (a) Model 2 scores along with newly projected data of the other meat type (b) Model 3 scores along with newly projected scores.

10. Compute the T^2 control chart for the models 2 and 3 having the newly projected scores as well. What can you conclude about the samples? Which deviate more: the chicken samples from the beef model, or vice-versa?

Figure 7. T^2 scores for models 2 and 3 with cross-model samples brought in.

11. Compute the contributions to T^2 scores. High contributions are for variables that differentiate from the model. If you are to discriminate between beef and chicken samples, which variables would you keep? Motivate your decision using the loading plots, biplots and previous results as well.

Figure 8. Contribution to T^2 scores for the models 2 and 3, for model data and test data (other mode's data).