Make a copy of this. Don’t insert here ☺ **Names**:

**Dataset description:** The dataset contains around 160 variables (measurements of proteins) of 30 healthy patients and 30 patients with diabetes. The modelling goal is to create a reduced list of variables able to distinguish between a healthy and a diabetes patient, via control charts.

1. *Handling missing values*. Fill in the table with top 5 variables and observations with missing values, and decide if you are going to drop out the row (it has too many missing values per row) or the column (it has too many missing variables per variable). Remember that you have noticing more variables than observations. Create a box plot that is going to show if the variables you are going to drop have differences in values between the two groups (“healthy” and “diabetes”).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| Variables (160) |  |  |  |  |  |
| Observations (30 + 30) |  |  |  |  |  |

Figure 1. Grouped box plot for the variables with missing values that are going to be excluded from analysis.

1. After creating a matrix with no missing values, divide the data in “healthy” and “diabetes” data. Center and scale the “healthy” data, and save the mean and standard deviation. Create a PCA model with the scaled healthy data, and plot the cumulative explained variance of the model as a function of the number of principal components.

Figure 2. Cumulative explained variance with the number of principal components.

1. Center and scale the “diabetes” data with the mean and standard deviation of the “healthy” data. Project the “diabetes” data into the “control” model. Create an extended scores matrix that includes the newly projected “diabetes” data into the “healthy” model. Plot the biplots of the extended scores for the first principal components, using different colors for the scores coming from the original model (“healthy” data) and the scores coming from the test data (“diabetes” data). Is there any principal component able to differentiate between the samples?

Figure 3. Biplots of the first principal components, with the test data ("diabetes" data) projected into the model.

1. Create T2 and SPEx charts for the models including different numbers of Principal Components. Project the test data (“diabetes” data) into these control charts. How many PC are needed for the entire test data to appear out of control bounds in the T2 chart? What about the SPEx chart?

Figure 4. T2 and SPEx Control charts for different numbers of principal components.

1. Compute the contributions of the test data set (“Diabetes” data) to the T2 and SPEx control charts, for the number of principal components chosen for each. Plot, in bar plots, these variable contributions, and select a value for the limit-contribution, if you are to select the variables that contribute the most to the chart values.

Figure 5. (a) Bar Plot of variable contributions to T2 scores of the test data. (b). Variable contributions to the SPEx values of the test data.

1. Select a limit for the variable contributions for both T2 and SPEx contributions, and make a list of the variables (variable indexes) above the limit. Construct 3 new PCA models (i) with only high-T2 contribution variables (ii) with common high-T2 and high-SPEx contribution variables (iii) with only high-SPEx contribution variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | **1. Initial** healthy model | 2. High-T2 (test) contribution vars | 3. High-T2-and-SPEx (test) contribution vars | 4. High-SPEx (test) contribution vars |
| No. Vars | all |  |  |  |
| Var idx | all |  |  |  |

1. Compute the biplots of the first principal components (PC 1 vs. PC 2) for the new models, after you have projected the test data as well and extended the scores list including test data scores. Mark, in your biplots, with different colors the “healthy” and “diabetes” observation scores. Can you discriminate between them better in the new biplots?

Figure 6. (a) Initial Model Biplot (PC.1 vs. PC.2) (b) Model 2 Biplot (PC.1 vs. PC.2) (c) Model 3 Biplot (PC.1 vs. PC.2) (d) Model 4 Biplot (PC.1 vs. PC.2)

1. Re-compute the control charts for PCs 1->3 in the three new models. Which is the least number of variables, which are the variables and what no. of PCs are needed for the model, to discriminate between “healthy” and “diabetes” patients?

Figure 7. (a) T2 chart M2, 1 PC (b) T2 chart M2, 2 PCs (c) T2 chart M2, 3 PCs

(d) T2 chart M3, 1 PC (e) T2 chart M3, 2 PCs (f) T2 chart M3, 3 PCs

*(g) T2 chart M4, 1 PC (h) T2 chart M4, 2 PCs (i) T2 chart M4, 3 PCs*

(j) Q chart M2, 1 PC (k) Q chart M2, 2 PCs (l) Q chart M2, 3 PCs

(m) Q chart M3, 1 PC (n) Q chart M3, 2 PCs (p) Q chart M3, 3 PCs

*(q) Q chart M4, 1 PC (r) Q chart M4, 2 PCs (s) Q chart M4, 3 PCs*