# EMSE6765

# Data Analysis Project: Principal Component Analysis

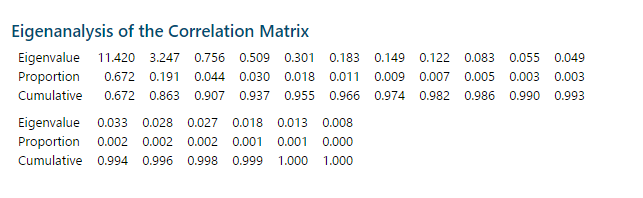
## Dataset Description

This dataset consists of Pea ID, Tenderometer, Dry matter, Dry matter after freeing, SucrosePercent, TotalGlucose1, TotalGlucose2, Flavour, Sweet, Fruity, Off-flavour, Mealiness, Hardness, Whiteness, Colour1, Colour2, Colour3, Skin data of 60 samples of peas. There are some substantial redundancy among dimensions, for example, between TotalGlucose1 and TotalGlucose2, which might lead to high levels of correlation and multicollinearity. Therefore, I am using PCA (Principal Component Analysis) to re-express this multivariate dataset, so that the first few dimensions will account for as much of the available information as possible. Also, in that case, each principal component is uncorrelated with all the others and hence independent pieces of information puzzle represented in the larger dataset.

This report details the Principal Component Analysis performed to assess and pick out the 10 peas that outperformed others.

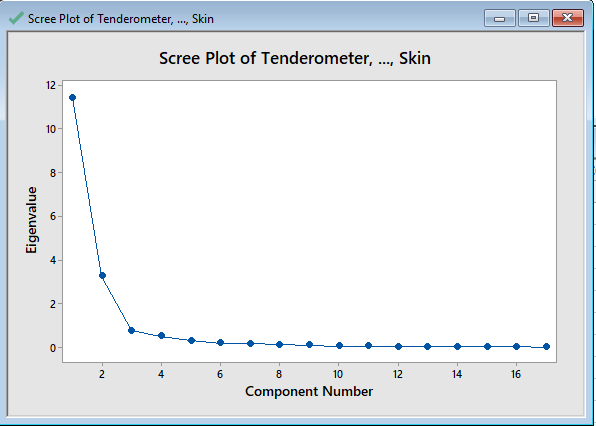
## Reduced Model Development

The first step is to perform the Principal Component Analysis. Eigenvectors and eigenvalues are computed in Minitab. The eigenvalues their percentage explanation of variance, and the cumulative contribution of each component are listed in Figure 1.



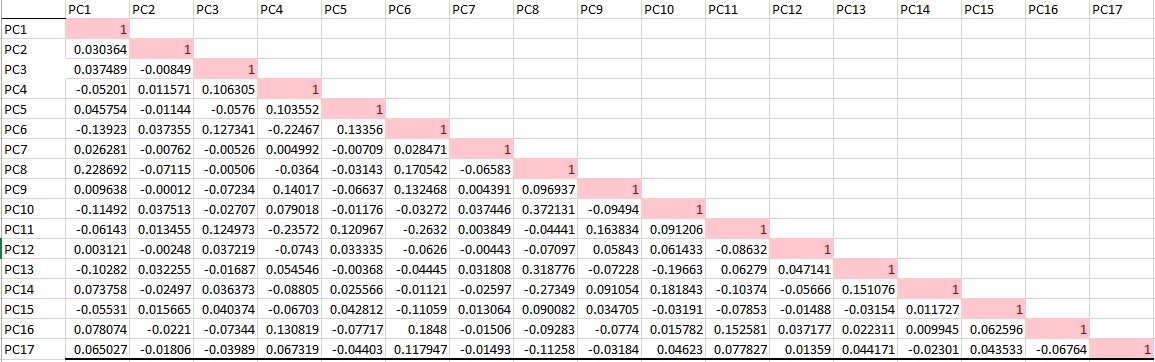
*Figure 1: Eigenvalues and Their Contribution*

The first step in the analysis is to determine how many components to be included in a reduced model. Based on Kaiser’s rule, the first 2 components are both greater than 1, so that they are pretty good and surely retained. The second approach is to check the scree plot shown as Figure 2.



*Figure 2: Scree Plot of PCA*

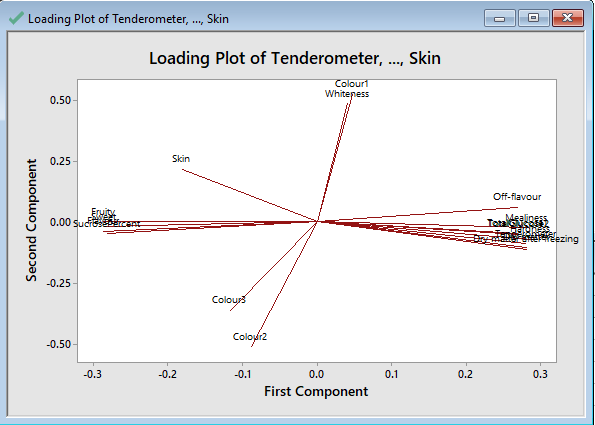
We can see that starting from the fifth point, the line starts to fall less steeply. Therefore from the results shown in the scree plot, I decided to retain the first four components. From the cumulative contribution shown in Figure 1, we can see that the first four components account for 93.7% of the total variance in the dataset. Also, from the correlation table of Principal Components shown as Figure 3, we can see that there’s no strong correlation between each two principal components. Therefore, I recommend selecting a reduced model with four components. The first four principal components account for 93.7% of the variance so it’s still good fit.



*Figure 3: Correlation Table of Principal Components*

## Data Assessment

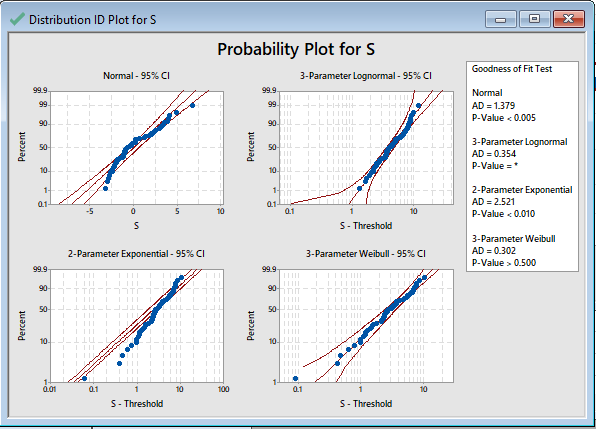
This analysis is performed using the loading plot provided as Figure 4. The first two components account for 86.3% of the total variation. The loading plot shows a number of groupings of pea characteristics/features: The first group consists of characteristics SucrosePercent, Flavour, Fruity and Sweet. Another tightly clustered group consists of characteristics Off-flavour, Mealiness, TotalGlucose1, TotalGlucose2, Hardness, Tenderometer, Dry matter and Dry matter after freezing. The slightly loosely clustered groups are the group consists of Colour1 and whiteness and the group consists of Colour2 and Colour3. From what is shown in the loading plot, there’s no feature clustered with feature Skin. For the first component, the group consists of Off-flavour, Mealiness, TotalGlucose1, TotalGlucose2, Hardness, Tenderometer, Dry matter and Dry matter after freezing seems to have a large positive loading (almost 0.3). The group consists of features SucrosePercent, Flavour, Fruity and Sweet seems to have the strongest negative loading(almost -0.3). However, for the second component, the group consists of Colour1 and Whiteness seem to have the highest positive loading (almost 0.5) on the Second Component Axis among which Colour1 seems to have a larger influence than Whiteness while the group consists of Colour2 and Colour3 seems to have the strongest negative loading (between -0.5 and -0.25) on the Second Component Axis among which, Colour3 seems to have a bigger influence than Colour2. However, even though skin has no other features to cluster with it, it still accounts for a large loading itself (almost 0.25).



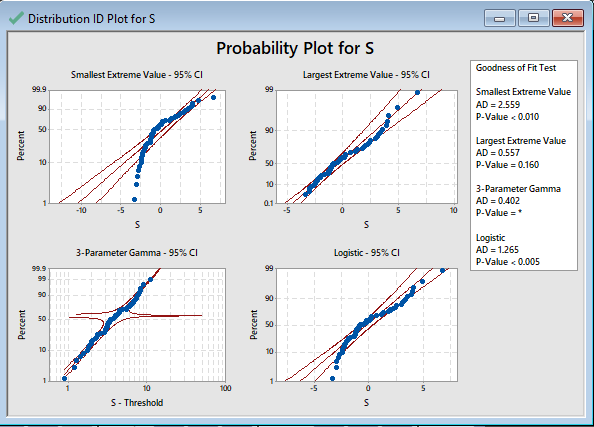
*Figure 4: Loading Plot of PCA*

## Pea Metric Design, Distribution Fitting and Pea Selection

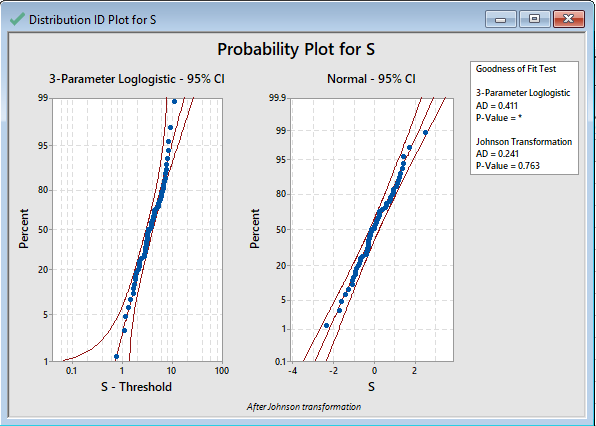
According to the score results from Principal Component Analysis, I formed a total score based on the proportion of the variance explained by the corresponding principal component. After that, I ran the individual distribution identification on the final score of the 60 peas, and the results turned out to be Figure 5, Figure 6 and Figure 7. From those figures, we can see that the distribution is more likely to be normal distribution after Johnson Transformation. And after selecting, we can see that the top 10% of the peas which outperformed others are pea 12, pea 15, pea 24, pea 27, pea 42 and pea 51, for they have the top 10% of the calculated score based on the principal components we chose and the principal components’ corresponding scores.



*Figure 5: Distribution Fitting 1*



*Figure 6: Distribution Fitting 2*



*Figure 6: Distribution Fitting 3*