FACHHOCHSCHULE
WIENER NEUSTADT
University of Applied Sciences – Austria

Demonstration of metabolomic data through PCA and PLS-DA Analysis

Chemometric

Dr. Ulrik Sundekilde

Student's name: Marjan Aziminezhad

Date: Feb. 2023

1. Introduction:

Genetic distance refers to the genetic divergence between species or between populations within a species. Smaller genetic distances indicate that the populations have more similar genes, which indicates they are closely related; they have a recent common ancestor, or recent interbreeding has taken place. In order to obtain the best results, checking for incompleteness of data and bootstrap-resampling to assess the stability of tree topologies should be considered.

1.1 Purpose

To demonstrate proficiency in performing PCA and PLS-DA on metabolomics data, handling outliers and
interpreting results.

1.2 Problem

To test if it is possible to develop models capable of differentiating between milk samples from dairy cows
in traditional and organic farming or from farms delivering milk to specific dairies. Milk samples are collected using automated milking in the normal routine at the farm.
Milk samples were analyzed for metabolites using 1H NMR spectroscopy.

1.3 Data

The dimension of the data structure is 89x34.
The first row is metabolite id.
The first column is a sample id, 2nd dairy and 3rd organic/conventional farming. Columns 4-34 are metabolites.
File name: dataforexamproject.xlsx

2. Methods

2.1 PCA:

a. Describe steps taken to reach the final PCA model

The initial stage of the analysis process involves preparing the data for analysis. This necessitates the cleaning and transformation of the data, as required. To identify the optimal model that best describes our data, we employed Principal Component Analysis (PCA) using the Ropls package.

It is crucial to examine the results of the PCA analysis. This involves scaling and centering the data by passing different scaling models to the `scaleC` function and observing the scores plot to determine the maximum variance in our data. Additionally, the explained variance of each component must be evaluated, starting with 10 components in this case, which explained 81% of the data variance. Subsequently, scree plots are used to determine the number of components to retain, and a biplot is created to visualize the relationships between variables and components.

It is important to identify outliers and have an overall view of the results to reduce residuals. The original data can be transformed into a new coordinate system defined by the principal components by multiplying the original data matrix by the matrix of loadings.

It is advisable to compare the first three principal components as they have the highest loading values to attain the best results.

b. Interpret the final PCA model

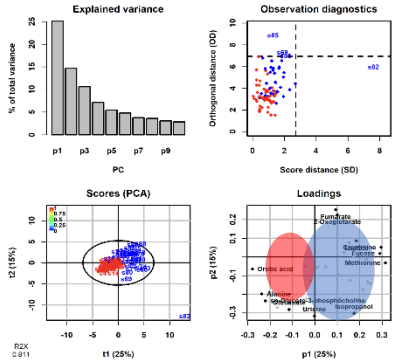

Figure 1 PCA of farming data using Standard scaling wit

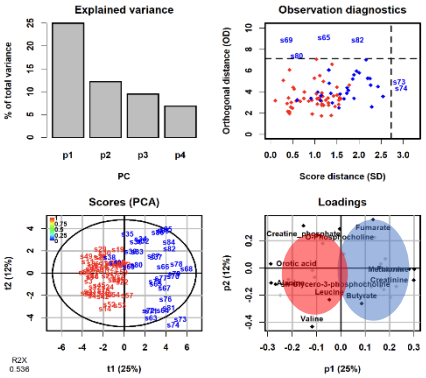

Figure 2 PCA of four farming componants data using Standard scaling without outliers

In PCA, a scatter plot with a more dispersed distribution of scores provides a better visual representation of the data and its variability. This is because PCA is a technique that reduces the dimensionality of the data while preserving the most crucial information by creating new variables, known as principal components, which explain the maximum variance in the original data.
The data represented by each loading in the variance diagram (scree plot) reduces as expected. The explained variance of each principal component is depicted in the graph, where the principal components with high explained variance capture a significant proportion of the variation in the data. The presence of a large proportion of variance explained by the first few principal components suggests that the data has a straightforward structure.
Figure 1 shows a scatter plot with tight scores, which may not capture the essential relationships in the data effectively. Additionally, the observation of S82 as a potential outlier, due to its substantial score distance from all other variables, is possible. To attain better results, it may be necessary to remove S82.
In contrast, Figure 2 depicts the principal components with a more dispersed distribution of scores, indicating that a greater proportion of the total variance in the data is captured. This provides a more comprehensive view of the data's structure, and thus, standard scaling is preferred in this case.
The loading plots correspond positionally to the score plots, with red scores having a higher load on the metabolic variables on the left side of the loading plot and vice versa.(not conveniently because of small overlapping) As an example, S14 has a higher amount of valine compared to all other variables, with a value of 0.02618262.
In conclusion, the amount of variance in the data explained by each principal component is displayed in the graph. The principal components with high explained variance capture a significant proportion of the variation in the data, and the first few principal components explain a large proportion of the variance, indicating that the data has a straightforward structure.

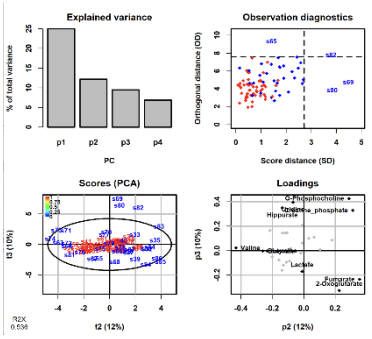

Figure 3 PCA of Farming data using 4 components and comparing P2 with P3

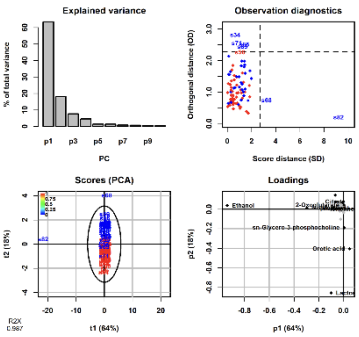

Figure 4 PCA of Farming data using Pareto scale.

In Figure 3 a tight positioning of variables prevents to have a good data analysis the same as choosing the wrong scaling method in figure 4 which leads to a great explained variance score in the the first loading but still not a optimal score plot .

c. Can we use the model to classify dairies or organic/conventional farming?

PCA can be used for unsupervised classification by grouping data into clusters based on similarities. With only two components, the grouping may be clearer, but for more components, further analysis with techniques such as PLS or PLS-DA may be necessary for improved results.

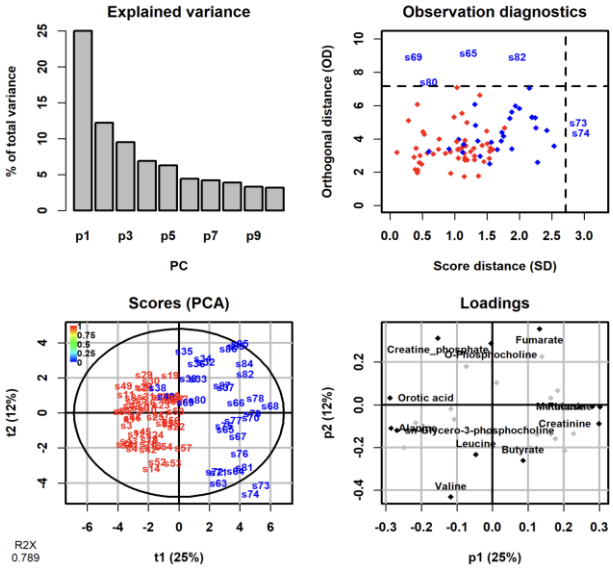

Figure 5 PCA of farming data -standard scaling

In Figure 5, the PCA scores plot for t1 shows a grouping of the data, but with some overlap. This overlap makes it difficult to interpret the data correctly, as the relationships between the data points are not clearly defined. This indicates that further analysis may be required to fully understand the structure of the data.

d. Can we identify differences in milk metabolome due to dairies or farming method?

The PCA plot displays the relationships between variables in a reduced number of dimensions, typically 2 or 3, which can make it easier to visualize differences in the data. PCA would not be enough in this level to provide information about the differences and to identify them.

2.2 PLS-DA

a. Describe steps taken to reach the final PLS-DA models

Load the required libraries: To start, you will need to load the ropls library and any other necessary libraries in R.

Prepare the data: Ensure that the data is properly formatted, including converting any non-numeric variables to factors and removing any missing values or outliers.

Split the data: Split the data into training and test sets(here crossvalI = 7). The training set will be used to fit the model, while the test set will be used to evaluate the performance of the model.

Preprocess the data: Perform any necessary preprocessing steps on the data, such as scaling, centering, or imputing missing values.

Fit the PLS-DA model: Use the opls function from the ropls library to fit the PLS-DA model on the training data. The function takes several arguments, including the response variable, the predictor variables, and the number of latent variables to use in the model.

Cross-validate the model: Use cross-validation to assess the performance of the model and tune any necessary parameters, such as the number of latent variables.

Evaluate the model: Use the test set to evaluate the performance of the final model, including accuracy, precision, recall, and F1 score.

Plot the results: Use visualization techniques, such as a score plot or a loadings plot, to visually assess the performance of the model and understand the relationships between the variables.

Interpreting the results: Interpret the results of the final model, including the variable importance, the contribution of each latent variable, and the relationships between the variables.

b. Interpret the final PLS-DA models

After cross validation, it is important to consider the maximum number of components for the bar chart to show the best result. R may choose 2, but the R2Y and Q2 table can be used to determine the maximum number of components. The aim is to avoid a decrease in explained Y variance. In this case, the maximum number of components for Dairy data has been chosen as 4 (predI=4) as it is represented in Table 1 and for farming data as 2(predI=2).

Table 1 Variance calculation of components

```
> examdairy.pls@modelDF
       R2X R2X(cum)     R2Y R2Y(cum)      Q2  Q2(cum) Signif. Iter.
p1  0.2540    0.254 0.21300    0.213  0.2050  0.20500       R1    13
p2  0.1020    0.356 0.13000    0.343  0.1150  0.29700       R1    29
p3  0.0517    0.408 0.12800    0.471  0.0344  0.32100       NS    40
p4  0.0610    0.469 0.08000    0.551  0.0167  0.33200       NS    14
p5  0.0457    0.515 0.03770    0.589 -0.1370  0.24100       NS    77
p6  0.0560    0.571 0.02360    0.612 -0.0864  0.17500       NS    48
p7  0.0538    0.625 0.01410    0.627 -0.1010  0.09170       NS    53
p8  0.0391    0.664 0.01230    0.639 -0.1120 -0.00977       NS   113
p9  0.0423    0.706 0.01130    0.650 -0.0876 -0.09830       NS    47
p10 0.0312    0.737 0.00921    0.659 -0.0648 -0.16900       N4    31
```

In Figures 6 and 7, the randomized validated Y is below zero and the model Y is on the line, which suggests that the model fits well. This conclusion is supported by the low pQ2 value, which indicates that the model has a good predictive ability. It is also important to identify and remove outliers to reduce bias in the analysis. Outliers can have a significant impact on the results, and removing them can improve the accuracy and reliability of the findings.

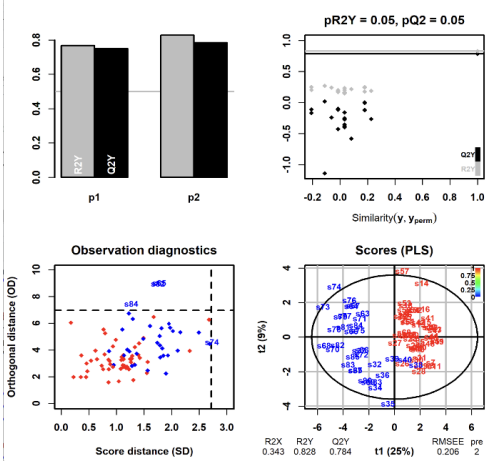

Figure 6 PLS of Farming data

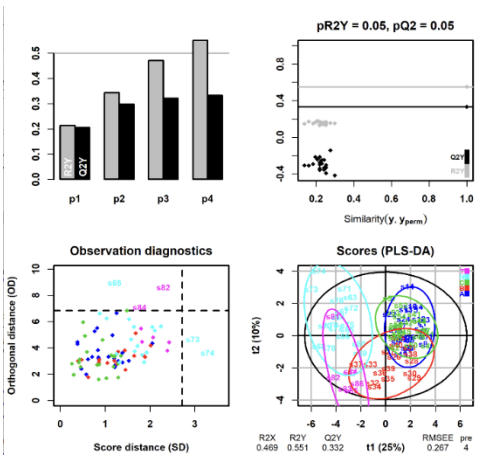

Figure 7 PLS of Dairy data

Table 2 PLS after cross validation for both Farming and Dairy data

```
> examdairy.pls <- opls(exams82.pls[,c(4:30)], exams82.pls$Dairy,  predI=4, crossvalI = 7, scaleC =
'standard', permI = 20)
PLS-DA
87 samples x 27 variables and 1 response
standard scaling of predictors and response(s)
      R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
Total    0.469    0.551   0.332 0.267   4   0 0.05 0.05
> examfarming.pls <- opls(exams82.pls[,c(4:30)], exams82.pls$Farming, predI=2, crossvalI = 7, scaleC
= 'standard', permI = 20)
PLS
87 samples x 27 variables and 1 response
standard scaling of predictors and response(s)
      R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
Total    0.343    0.828   0.784 0.206   2   0 0.05 0.05
```

The first component (Dairy) explains 0.33 of the variation in Y and The first component (Farming) explains 0.78 of the variation in Y.

A low Q2 (cumulative) value after cross-validation of data in a PLS similarity plot indicates that the model is not explaining the variance in the data well. This can be due to overfitting, insufficient latent variables, unbalanced data, outliers, or poor model selection. To improve the Q2 (cumulative) value, it may be necessary to modify the data or choose a different statistical method.

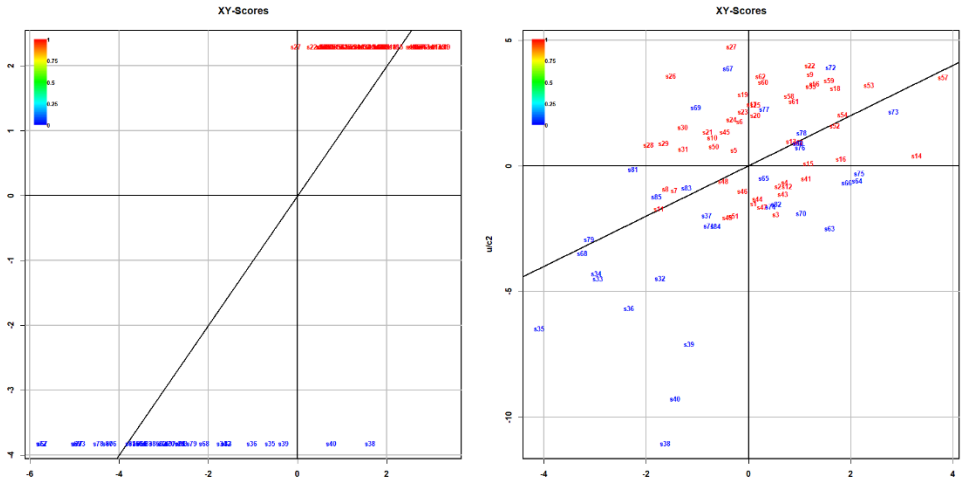

Figure 8 XY-Scores plot from Farming data

On the left XY-scores plot, we can observe two irrelevant variables. However, the P2 plot on the right shows good spread components with a linear correlation. We cannot suggest S38 and S40 as outliers based on their position in the PCA score plot as observed previously.

c. Can we use the model to classify dairies or organic/conventional farming?

There could be a classification of groups in both data.(see score plots)

d. Can we identify differences in milk metabolome due to dairies or farming method?

PCA and PLS-DA are dimension reduction techniques allow to visualize the relationships between variables in a reduced number of dimensions, typically 2 or 3. These plots can help identify patterns and differences in the data, but they are not conclusive in themselves. Still overlapped data could lead into wrong conclusions about differences .

The idea is that our data is supervised when we use PLS or PLS-DA methods which mean we have information about groups of data.

e. Is the interpretation the same comparing with the PCA?

The interpretation of PLS results focuses on the contribution of the predictor variables to the response variable, whereas the interpretation of PCA results focuses on the contribution of the original variables to the principal components. In other words, PLS interpretation is more focused on prediction, while PCA interpretation is more focused on pattern identification.

Additionally, to identify differences, we would have to examine the weights in the loading vectors, as highly correlated variables would appear close together in the loading plots of all dimensions. In PCA, the objective is to calculate each latent variable so that it best explains the available variance in X. In contrast, in PLS, the differences between variables are already given to the model. The goal of PLS is to find the scores that have the highest covariance. To do this, we need to examine the regression coefficients as a function of all the latent variables in the model and identify the highest ones. These high-regression coefficients can then be used to predict new samples. By doing so, it may be possible to reduce the number of variables used by keeping only those with the highest regression coefficients.

f. what is the advantage of using PLS-DA and what do we need to ensure when using PLS-DA?

When using PLS-DA, it is important to ensure that the data used is pre-processed properly, as the method is sensitive to outliers and the presence of irrelevant variables. Additionally, PLS-DA is a supervised learning method, meaning that the groups to be discriminated must be known beforehand, so it is important to have a well-defined and balanced class label representation in the data set.

2.3 PLS-DA

a. Describe steps taken to reach the final PLS-DA models

To construct a PLS-DA model using the "ropls" library in R, several steps must be followed,

- The data is pre-processed to remove any noise and to normalize the variables.
- The model builds a predictive relationship between the predictors (input variables) and the response (class label).
- The model uses partial least squares regression to identify the variables that are most predictive of the class label.

It also uses these predictive variables to classify the samples into one of two or more classes. including data preparation, model selection, model fitting, model evaluation, model refinement, result interpretation, and application to new data. In this context, it is necessary to first perform a PCA on the data. This involves cleaning and transforming the data, finding the optimal model that describes the data, analyzing the results, and reducing outliers to reduce bias. The PCA results provide important information for building the PLS-DA model, as the PLS-DA algorithm requires a reduced and transformed version of the data.

b. Interpret the final PLS-DA models

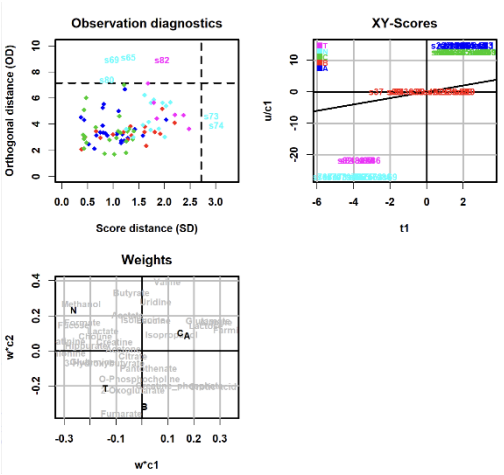

Figure 7 PLS-DA of Dairy

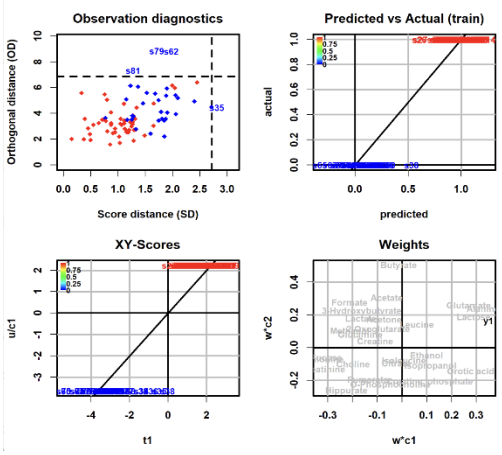

Figure 8 PLS-DA of Farming

The XY-scores plot of Figure 7 suggests the separation of the data into five linear groups, with each group being surrounded by the most contributing original variables (e.g. lactose, glutamate, and alanine for Figure 8). Although partial least squares-discriminant analysis (PLS-DA) is typically considered to be effective when dealing with two to four classes, this may be a contributing factor to the observation of a clear separation in the red group, but not in the other four groups, despite their close proximity. In contrast, a remarkable separation of two classes can be seen in the XY-scores plot of Figure 8.

c. Can we use the model to classify dairies or organic/conventional farming?

Yes, In PLS-DA, a scatter plot is used to visualize the separation between the classes in the data. The plot represents the samples in a two-dimensional space where the first dimension is the first PLS component and the second dimension is the second PLS component. Samples from different classes are colored differently and it can be seen how well the classes are separated in this two-dimensional space.

Additionally, a "loading plot" can be used to visualize the contribution of the variables to the separation between the classes. In this plot, the variables are represented as vectors and the angle between the vectors indicates the correlation between the variables and the classes.

By examining these plots, one can gain insight into the variables that are most important in separating the classes and the relationships between the variables and the classes. This information can then be used to improve the PLS-DA model or to develop new models for classifying new data.

d. Can we identify differences in milk metabolome due to dairies or farming method?

The model can be used to identify differences in the milk metabolome due to dairies or farming method. The model coefficients and scores on the latent variables can be used to identify the most important metabolites that are contributing to the class separation. This can provide insight into the differences in the metabolomes of the different dairies or farming methods.

e. Is the interpretation the same comparing with the PCA?

PCA and PLS-DA are two different techniques with different goals. PCA is an unsupervised method that aims to find the directions of maximum variance in the data(prediction), while PLS-DA is a supervised method that aims to separate the classes based on the predictors and response variables. As a result, the interpretation of the two methods can be different. In PCA, the principal components are the directions of maximum variance in the data, and samples can be visualized in a score plot based on

their scores on the principal components. In PLS-DA, the latent variables are optimized to separate the classes, and samples are visualized in a score plot based on their scores on the latent variables.

f. What is the advantage of using PLS-DA and what do we need to ensure when using PLS-DA?

Advantages: PLS-DA is a supervised method, which means it considers both the predictor variables and the response variables in the analysis. This makes it well suited for classification problems, where the goal is to predict a categorical response variable based on a set of predictor variables. PLS-DA is also able to handle highly correlated predictor variables, which is often a challenge in other classification techniques.

Considerations: PLS-DA is a linear method, which means it assumes a linear relationship between the predictors and response variables. If the relationship is non-linear, PLS-DA may not perform well. Additionally, PLS-DA is sensitive to the choice of number of latent variables, and too few or too many latent variables can negatively affect the performance of the model. It is also important to ensure that the data is pre-processed properly, such as scaling the data, to ensure that the model is not influenced by the scale of the variables. Finally, it is important to evaluate the performance of the model on an independent test set, to ensure that it generalizes well to new data.