# Partial Least Square - Discriminant Analysis (PLSDA) Summary:

PLS-DA is a chemometrics technique used to optimise separation between different groups of samples, which is accomplished by linking two data matrices X (i.e., raw data) and Y (i.e., groups, class membership etc.). The method is in fact an extension of PLS1 which handles single dependent continues variable whereas PLS2 (called PLS-DA) can handle multiple dependent categorical variables. This approach aims to maximize the covariance between the independent variables X (sample readings; that is to say the metabolomics data) and the corresponding dependent variable Y (classes, groups; that is to say the targets that one wants to predict) of highly multidimensional data by finding a linear subspace of the explanatory variables. This new subspace permits the prediction of the Y variable based on a reduced number of factors (PLS components, or what are also known as latent variables). These factors describe the behavior of dependent variables Y and they span the subspace onto which the independent variables X are projected.

The main advantage of this PLS-DA approach is the availability and handling of highly collinear and noisy data, which are very common outputs from metabolomics experiments. In addition, this provides several statistics such as loading weight, variable importance on projection (VIP) and regression coefficient that can be used to identify the most important variables. This technique provides a visual interpretation of complex datasets through a low-dimensional, easily interpretable scores plot that illustrates the separation between different groups. Comparison of loadings and scores plot supports investigations in terms of the relationship between important variables that can be specific to the group of interest.

### Input Summary:

Input Dataset: -> e.csv

Output Datasets and Files: -> plsda -> perm\_plot.svg -- vip\_plot.svg -- scree\_plot.svg -- loading\_plot.svg -- score\_plot.svg -- compound\_loadings.csv -- sample\_scores.csv

**- Treatment Group:** Genotype. The PLS-DA model will be performed to find a linear transformation on the compounds to discriminant the Genotype group.

**- Scaling Method:** Auto Scaling. Scaling methods are data pretreatment approaches that divide each variable by a factor, the scaling factor, which is different for each variable. They aim to adjust for the differences in fold differences between the different metabolites by converting the data into differences in concentration relative to the scaling factor. It is highly recommended for multivariate statistical analyses. More information is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1534033/.

### Result Summary:

The dataset was first scaled using Auto Scaling. Then the Partial Least Square - Discriminant Analysis (PLS-DA) was performed on the scaled dataset. When the number of predictive components equals to 8, the model achieves the highest Q2 score. In this model, the R2X (variances explained), R2Y, and Q2 (cum) is 4.8119, and 1, respectively. See Figure 1, Figure 2 and Figure 3 for more details.

Figure 1

Scores plot. It can be used to visually validate the fitness of the PLS-DA model. If the between-group variation is obvious, it indicates a goodness of fit. Samples colors/shapes/sizes with 95% confidence intervals can be added to visualize the sample clusters.

Figure 2

### Result Summary:

The dataset was first scaled using Auto Scaling. Then the Partial Least Square - Discriminant Analysis (PLS-DA) was performed on the scaled dataset. When the number of predictive components equals to 8, the model achieves the highest Q2 score. In this model, the R2X (variances explained), R2Y, and Q2 (cum) is 4.8119, 1 and 1, respectively. See Figure 1, Figure 2 and Figure 3 for more details.

Figure 1

Scores plot. It can be used to visually validate the fitness of the PLS-DA model. If the between-group variation is obvious, it indicates a goodness of fit. Samples colors/shapes/sizes with 95% confidence intervals can be added to visualize the sample clusters.

Figure 2

### Result Summary:

The dataset was first scaled using Auto Scaling. Then the Partial Least Square - Discriminant Analysis (PLS-DA) was performed on the scaled dataset. When the number of predictive components equals to 8, the model achieves the highest Q2 score. In this model, the R2X (variances explained), R2Y, and Q2 (cum) is 4.8119, 1 and 1, respectively. See Figure 1, Figure 2 and Figure 3 for more details.

Figure 1

Scores plot. It can be used to visually validate the fitness of the PLS-DA model. If the between-group variation is obvious, it indicates a goodness of fit. Samples colors/shapes/sizes with 95% confidence intervals can be added to visualize the sample clusters.

Figure 2

Loadings plot. It summarizes the linear relationship/similarity between the compounds. Compounds colors/shapes/sizes with 95% confidence intervals can be added afterwards to visualize the compound clusters. Together with the scores plot (Figure 1), loadings plot can help toidentify compounds contributing to between-group variability based on separations observed between groups in the scores plot.

Figure 3

Screes plot. Commonly, R2X and R2Y represent the fraction of variance of the X and Y matrix, respectively, and Q2 represents the predictive accuracy of the model, with cumulative (cum) values of R2X, R2Y and Q2 equating to ~1 indicating an effective model. Please note, when the number of components increase, the total sum of the R2X and R2Y increase, but not necessarily the Q2 as it is cross-validated R2Y. Generally, the higher the Q2 the better the model is. If R2X and R2Y is high but Q2 is too low, it means a crisis of overfitting.

This model achieves a Q2 of 1 at the 8 component.

Figure 4

Vip score plot. The VIP (Variable Importance in Projection) quantifies the contribution of a compound when building the model. Usually, a VIP score greater than one is considered important and (positively) affect classification between the groups.

On the right hand side of the vip score plot is a simple heatmap, indicating the changing direction of the compounds.

This model achieves a Q2 of 1 at the 8 component.

Figure 5

Permutation test plot. A permutation test can evaluate whether the PLS-DA classification in the designed groups is significantly better than any other random classification in arbitrary groups.

In our case, 100 permutations was performed on 8 components (which achieved the highest Q2 score). The p-value of the R2Y is 1, while the Q2 (cum) is 0.07. At least one of the p-values of R2Y and Q2 (cum) less than 0.05 indicates a valid model. Otherwise, it is hard to justify whether the designed dataset is different from a random arbitrarty dataset.

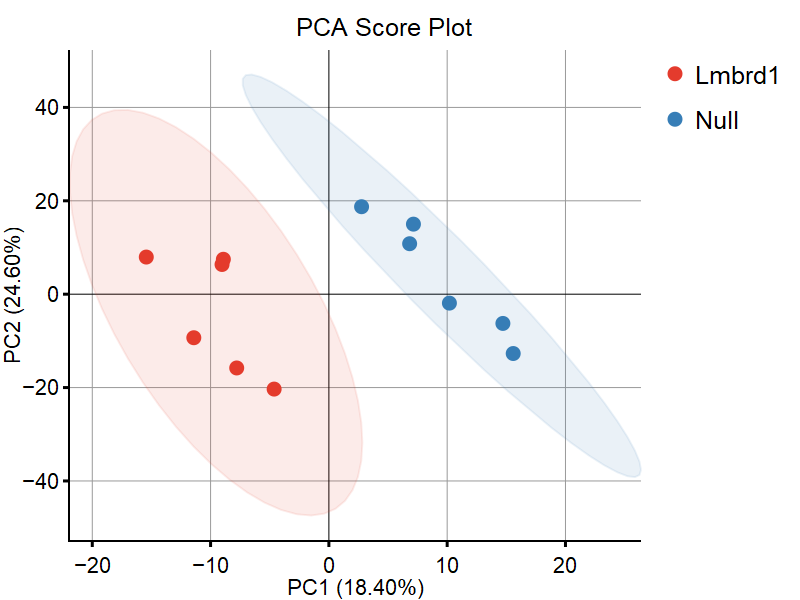


Figure 1: PLS-DA Scores Plot.

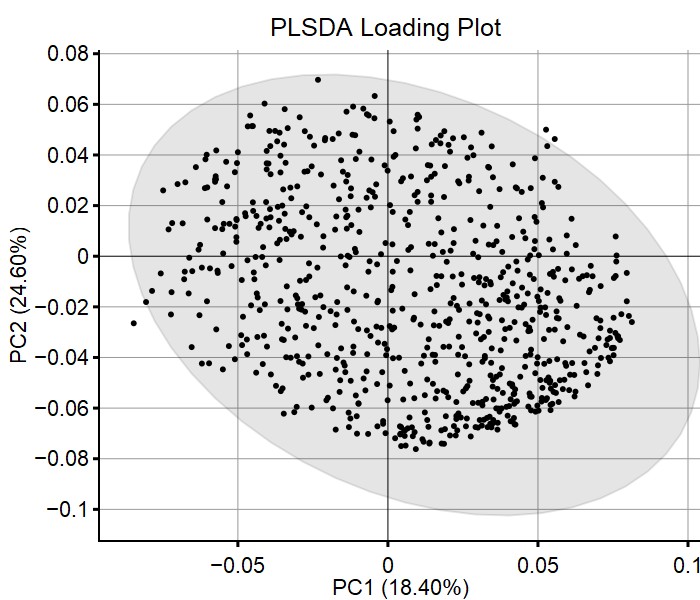


Figure 2: PLS-DA Loadings Plot.

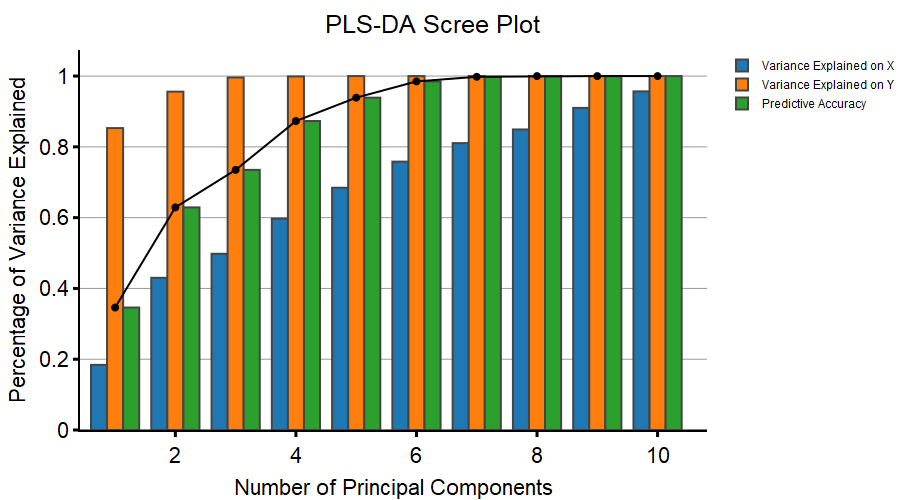


Figure 3: PLS-DA Scree Plot.

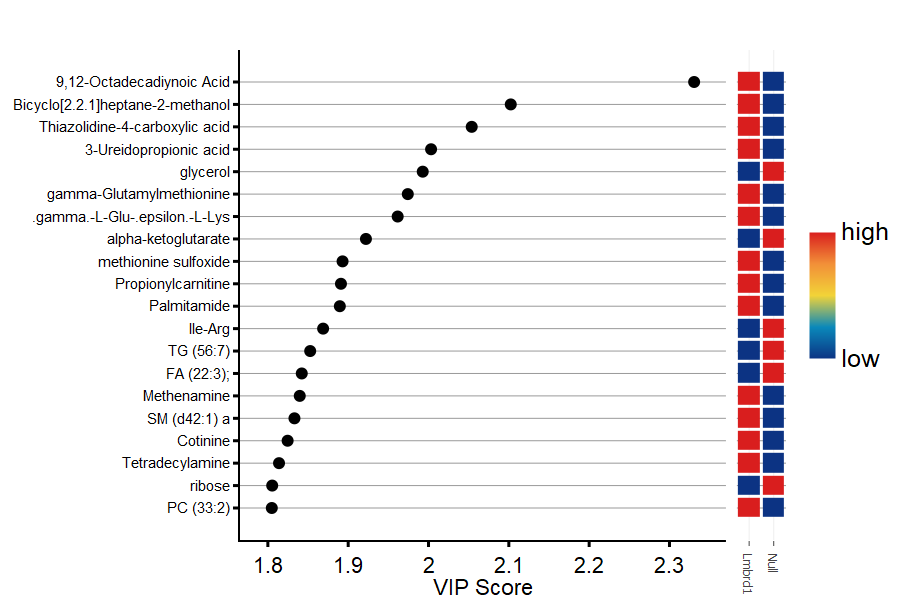


Figure 4: PLS-DA VIP Score Plot.

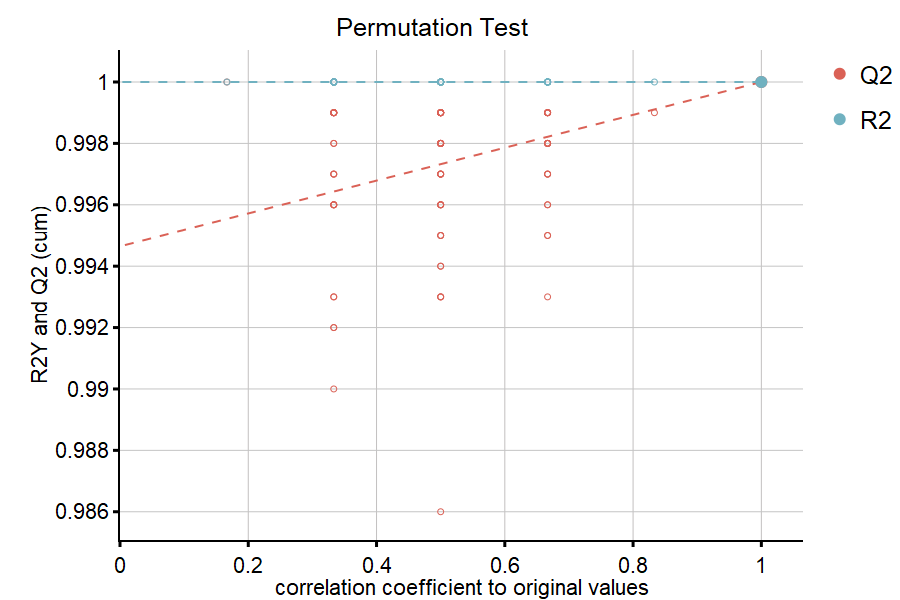


Figure 5: PLS-DA Permutation Plot.