

Metabolomic Data Analysis with MetaboAnalyst 3.0

User ID: guest3557652992311388261

January 1, 2017

1 Data Processing and Normalization

1.1 Reading and Processing the Raw Data

MetaboAnalyst accepts a variety of data types generated in metabolomic studies, including compound concentration data, binned NMR/MS spectra data, NMR/MS peak list data, as well as MS spectra (NetCDF, mzXML, mzDATA). Users need to specify the data types when uploading their data in order for MetaboAnalyst to select the correct algorithm to process them. Table 1 summarizes the result of the data processing steps.

1.1.1 Reading Binned Spectral Data

The binned spectra data should be uploaded in comma separated values (.csv) format. Samples can be in rows or columns, with class labels immediately following the sample IDs.

Samples are in rows and features in columns The uploaded file is in comma separated values (.csv) format. The uploaded data file contains 10 (samples) by 1018 (spectra bins) data matrix.

1.1.2 Data Integrity Check

Before data analysis, a data integrity check is performed to make sure that all the necessary information has been collected. The class labels must be present and contain only two classes. If samples are paired, the class label must be from $-n/2$ to -1 for one group, and 1 to $n/2$ for the other group (n is the sample number and must be an even number). Class labels with same absolute value are assumed to be pairs. Compound concentration or peak intensity values should all be non-negative numbers. By default, all missing values, zeros and negative values will be replaced by the half of the minimum positive value found within the data (see next section)

1.1.3 Missing value imputations

Too many zeroes or missing values will cause difficulties for downstream analysis. MetaboAnalyst offers several different methods for this purpose. The default method replaces all the missing and zero values with a small values (the half of the minimum positive values in the original data) assuming to be the detection limit. The assumption of this approach is that most missing values are caused by low abundance metabolites (i.e. below the detection limit). In addition, since zero values may cause problem for data normalization (i.e. \log), they are also replaced with this small value. User can also specify other methods, such as replace by mean/median, or use K-Nearest Neighbours, Probabilistic PCA (PPCA), Bayesian PCA (BPCA) method, Singular Value Decomposition (SVD) method to impute the missing values ¹. Please choose the one that is the most appropriate for your data.

¹Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. *pcMethods: a bioconductor package, providing PCA methods for incomplete data.*, Bioinformatics 2007 23(9):1164-1167

Zero or missing variables were replaced with a small value: 2.82e-08

1.1.4 Data Filtering

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step can usually improves the results. Data filter is strongly recommended for datasets with large number of variables (> 250) datasets contain much noise (i.e.chemometrics data). Filtering can usually improve your results².

For data with number of variables < 250 , this step will reduce 5% of variables; For variable number between 250 and 500, 10% of variables will be removed; For variable number bwtween 500 and 1000, 25% of variables will be removed; And 40% of variabed will be removed for data with over 1000 varaibles.

No data filtering was applied

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
c1	653	365	1018
c2	733	285	1018
c3	841	177	1018
c4	767	251	1018
c5	808	210	1018
e1	954	64	1018
e2	935	83	1018
e3	897	121	1018
e4	649	369	1018
e5	1005	13	1018

²Hackstadt AJ, Hess AM. *Filtering for increased power for microarray data analysis*, BMC Bioinformatics. 2009; 10: 11.

1.2 Data Normalization

The data is stored as a table with one sample per row and one variable (bin/peak/metabolite) per column. The normalization procedures implemented below are grouped into four categories. Sample specific normalization allows users to manually adjust concentrations based on biological inputs (i.e. volume, mass); row-wise normalization allows general-purpose adjustment for differences among samples; data transformation and scaling are two different approaches to make features more comparable. You can use one or combine both to achieve better results.

The normalization consists of the following options:

1. Sample specific normalization (i.e. normalize by dry weight, volume)
2. Row-wise procedures:
 - Normalization by the sum
 - Normalization by the sample median
 - Normalization by a reference sample (probabilistic quotient normalization)³
 - Normalization by a reference feature (i.e. creatinine, internal control)
3. Data transformation :
 - Generalized log transformation (glog 2)
 - Cube root transformation
4. Data scaling:
 - Unit scaling (mean-centered and divided by standard deviation of each variable)
 - Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)
 - Range scaling (mean-centered and divided by the value range of each variable)

Figure 1 shows the effects before and after normalization.

³Dieterle F, Ross A, Schlotterbeck G, Senn H. *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics*, 2006, Anal Chem 78 (13);4281 - 4290

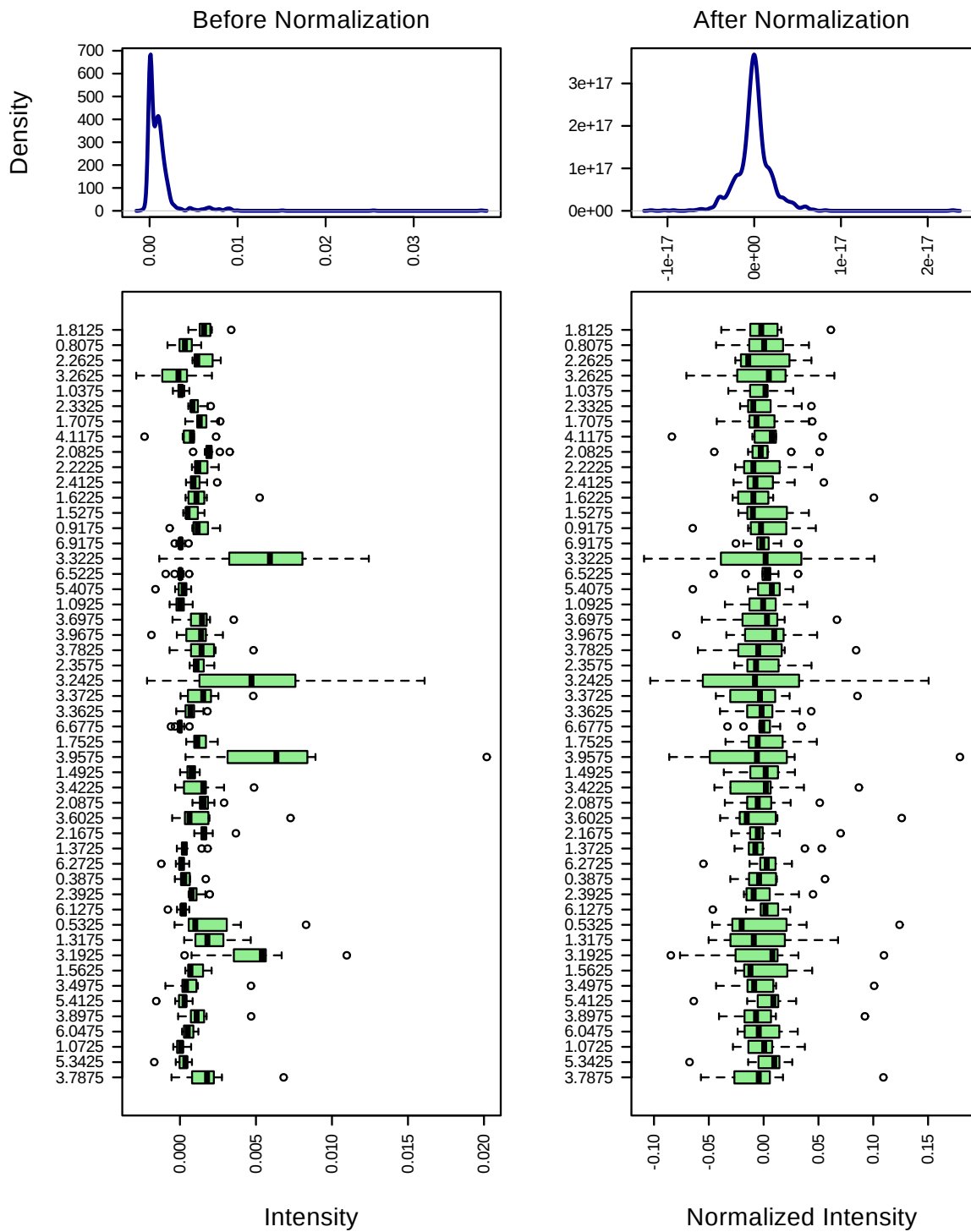


Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 50 features due to space limit. The density plots are based on all samples. Selected methods : Row-wise normalization: N/A; Data transformation: N/A; Data scaling: Pareto Scaling.

2 Statistical and Machine Learning Data Analysis

MetaboAnalyst offers a variety of methods commonly used in metabolomic data analyses. They include:

1. Univariate analysis methods:
 - Fold Change Analysis
 - T-tests
 - Volcano Plot
 - One-way ANOVA and post-hoc analysis
 - Correlation analysis
2. Multivariate analysis methods:
 - Principal Component Analysis (PCA)
 - Partial Least Squares - Discriminant Analysis (PLS-DA)
3. Robust Feature Selection Methods in microarray studies
 - Significance Analysis of Microarray (SAM)
 - Empirical Bayesian Analysis of Microarray (EBAM)
4. Clustering Analysis
 - Hierarchical Clustering
 - Dendrogram
 - Heatmap
 - Partitional Clustering
 - K-means Clustering
 - Self-Organizing Map (SOM)
5. Supervised Classification and Feature Selection methods
 - Random Forest
 - Support Vector Machine (SVM)

Please note: some advanced methods are available only for two-group sample analysis.

2.1 Partial Least Squares - Discriminant Analysis (PLS-DA)

PLS is a supervised method that uses multivariate regression techniques to extract via linear combination of original variables (X) the information that can predict the class membership (Y). The PLS regression is performed using the `pls` function provided by R `pls` package⁴. The classification and cross-validation are performed using the corresponding wrapper function offered by the `caret` package⁵.

To assess the significance of class discrimination, a permutation test was performed. In each permutation, a PLS-DA model was built between the data (X) and the permuted class labels (Y) using the optimal number of components determined by cross validation for the model based on the original class assignment. `MetaboAnalyst` supports two types of test statistics for measuring the class discrimination. The first one is based on prediction accuracy during training. The second one is separation distance based on the ratio of the between group sum of the squares and the within group sum of squares (B/W-ratio). If the observed test statistic is part of the distribution based on the permuted class assignments, the class discrimination cannot be considered significant from a statistical point of view.⁶

There are two variable importance measures in PLS-DA. The first, Variable Importance in Projection (VIP) is a weighted sum of squares of the PLS loadings taking into account the amount of explained Y-variation in each dimension. Please note, VIP scores are calculated for each components. When more than components are used to calculate the feature importance, the average of the VIP scores are used. The other importance measure is based on the weighted sum of PLS-regression. The weights are a function of the reduction of the sums of squares across the number of PLS components. Please note, for multiple-group (more than two) analysis, the same number of predictors will be built for each group. Therefore, the coefficient of each feature will be different depending on which group you want to predict. The average of the feature coefficients are used to indicate the overall coefficient-based importance.

Figure 2 shows the overview of scores plots; Figure 3 shows the 2-D scores plot between selected components; Figure 4 shows the 3-D scores plot between selected components; Figure 5 shows the loading plot between the selected components; Figure 6 shows the classification performance with different number of components; Figure 7 shows the results of permutation test for model validation; Figure 8 shows important features identified by PLS-DA.

⁴Ron Wehrens and Bjorn-Helge Mevik. *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*, 2007, R package version 2.1-0

⁵Max Kuhn. Contributions from Jed Wing and Steve Weston and Andre Williams. *caret: Classification and Regression Training*, 2008, R package version 3.45

⁶Bijlsma et al. *Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation*, *Anal Chem.* 2006, 78 567 - 574

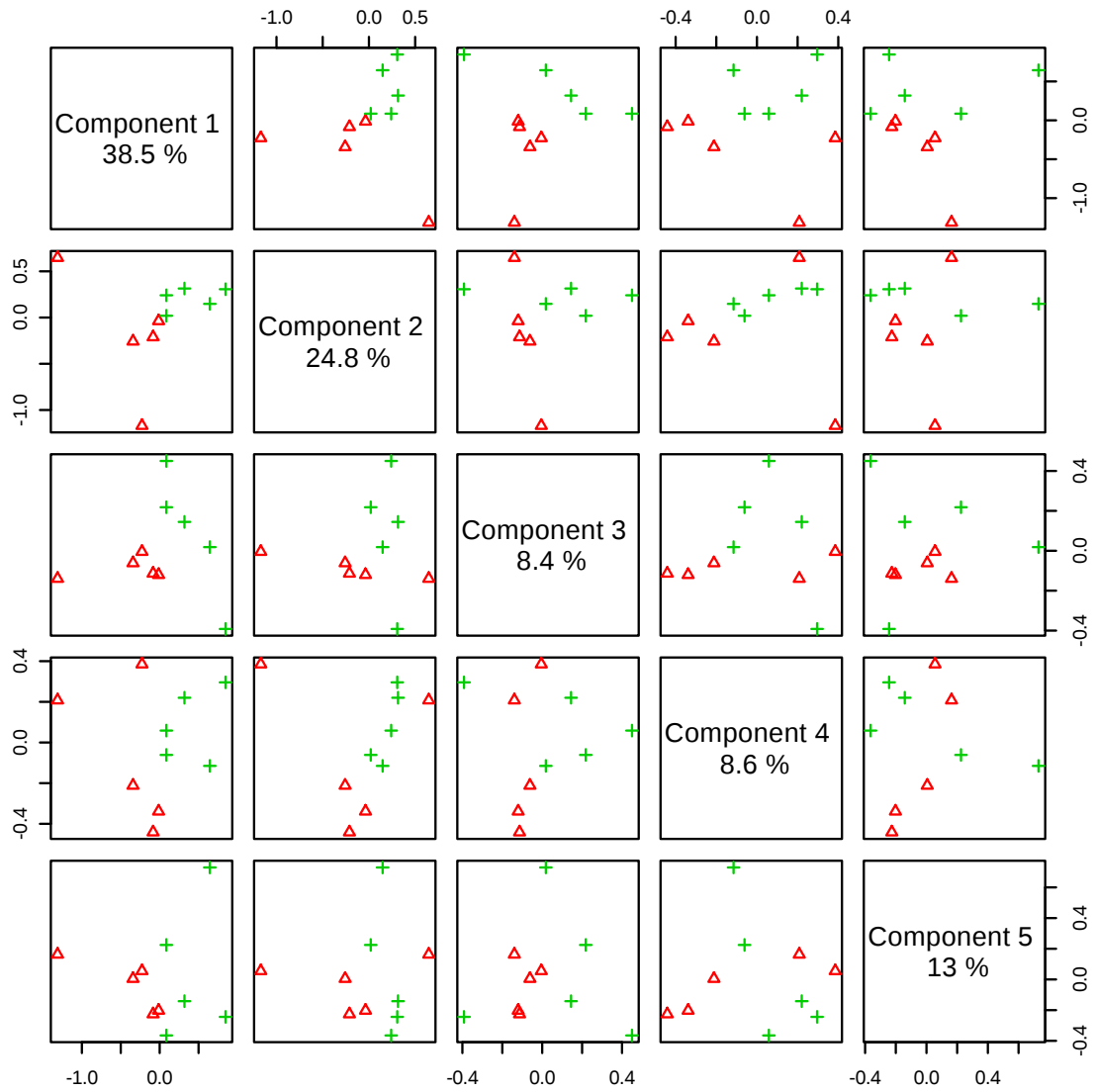


Figure 2: Pairwise scores plots between the selected components. The explained variance of each component is shown in the corresponding diagonal cell.

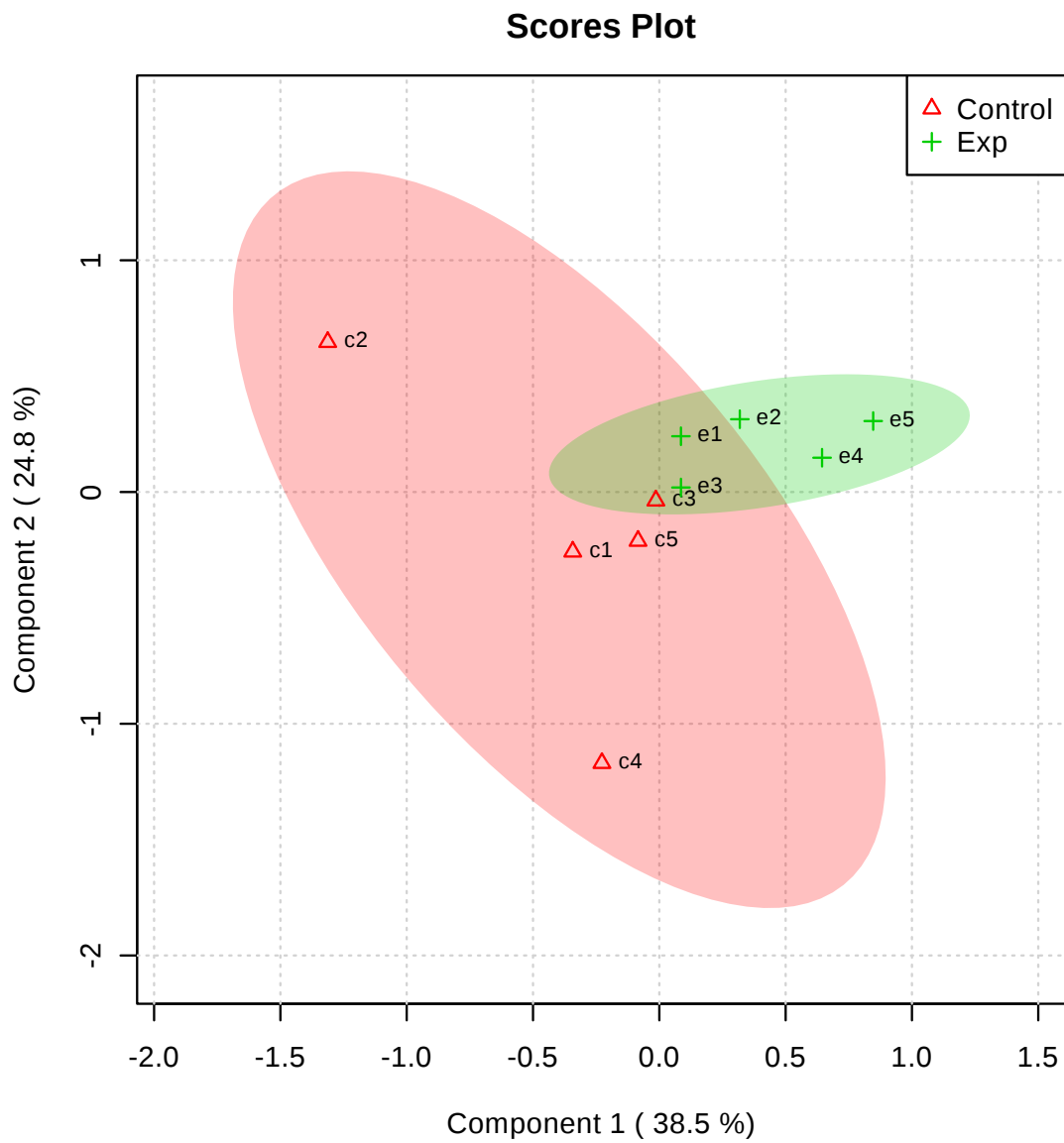


Figure 3: Scores plot between the selected PCs. The explained variances are shown in brackets.

Figure 4: 3D scores plot between the selected PCs. The explained variances are shown in brackets.

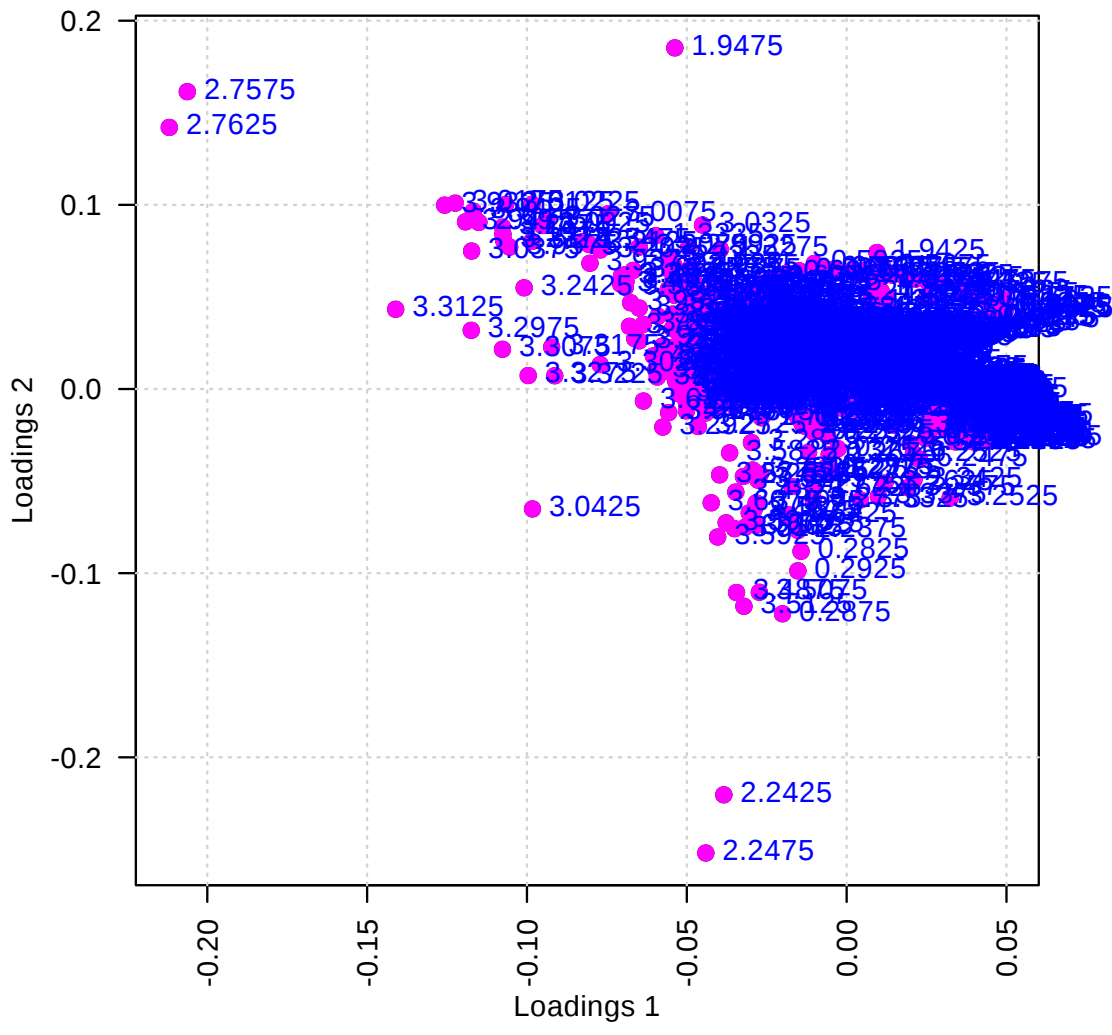


Figure 5: Loadings plot between the selected PCs.

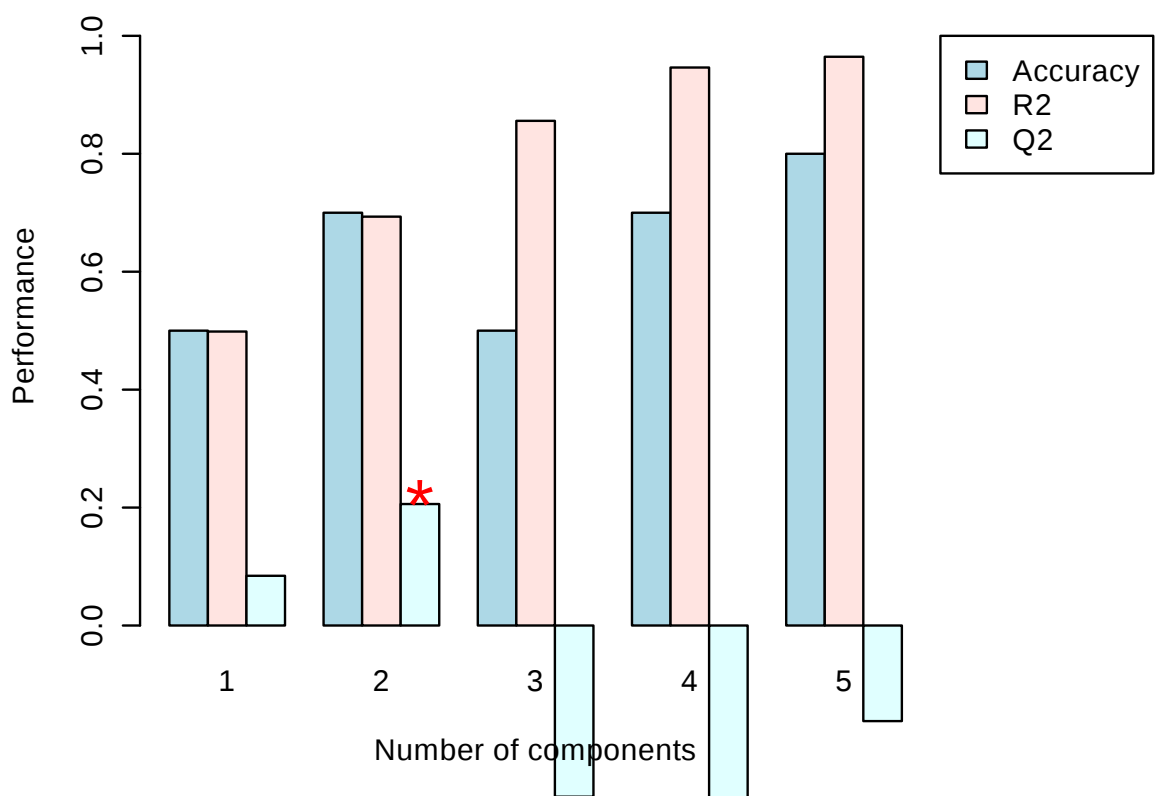


Figure 6: PLS-DA classification using different number of components. The red circle indicates the best classifier.

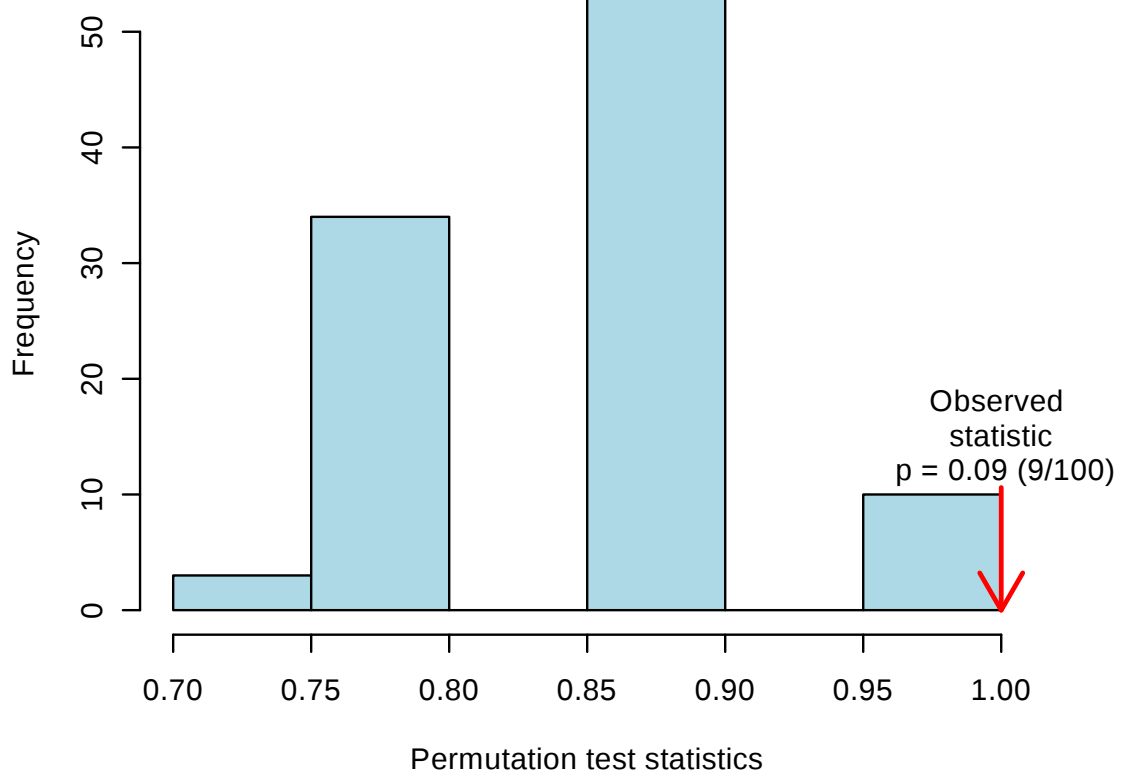


Figure 7: PLS-DA model validation by permutation tests based on prediction accuracy. The p value based on permutation is $p = 0.09$ (9/100).

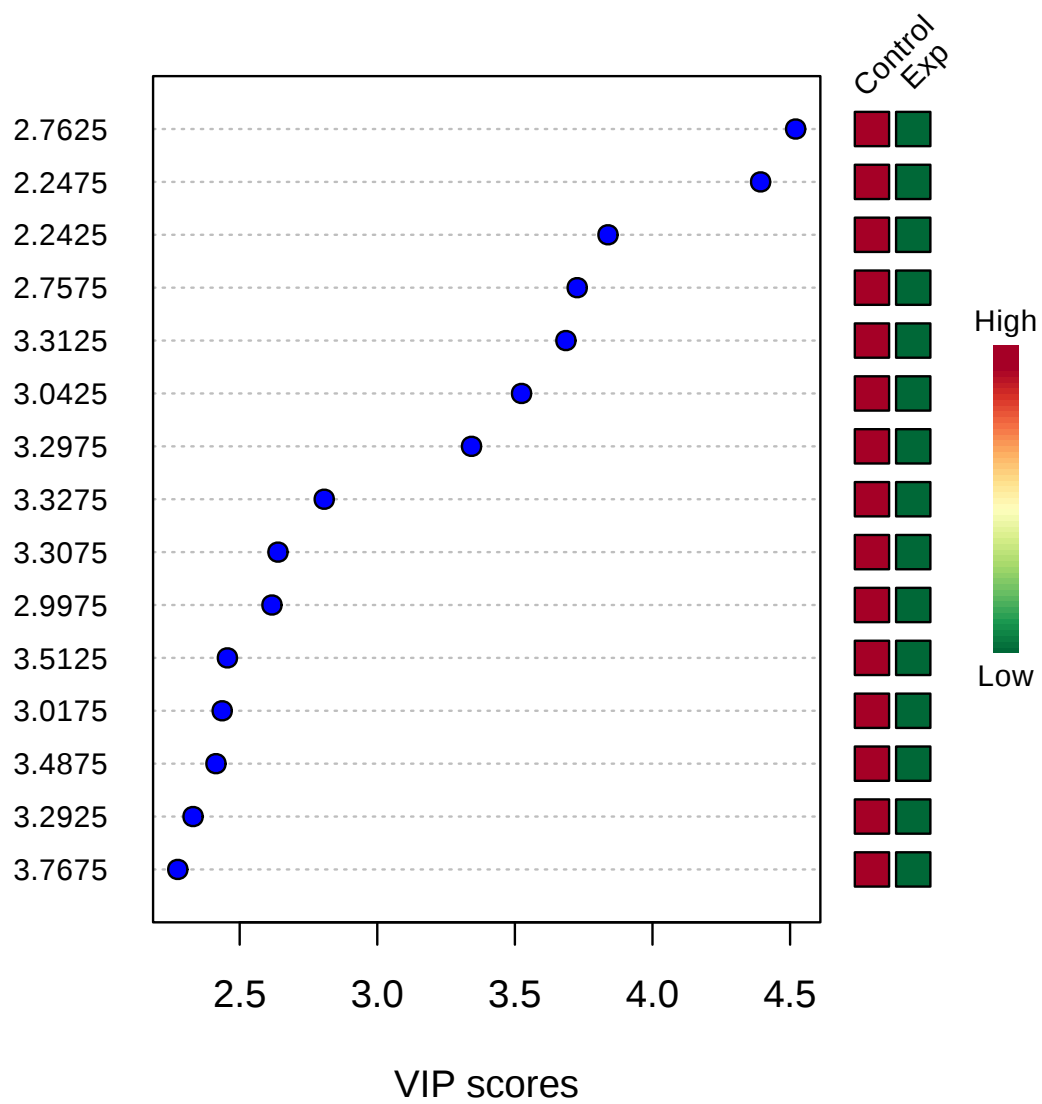


Figure 8: Important features identified by PLS-DA. The colored boxes on the right indicate the relative concentrations of the corresponding metabolite in each group under study.

3 Other Features

Please be advised that some more recent functions (Orthogonal PLS-DA, Power analysis, Biomarker analysis, Joint pathway analysis) with their corresponding results are not included in this report. For compound concentration data, users can perform metabolite set enrichment analysis and metabolic pathway analysis.

The report was generated on Sun Jan 1 11:53:49 2017 with R version 3.3.2 (2016-10-31).