POMAShiny Help Manual

Contents

0.0.1 Upload Data Panel

In this panel users can upload their data to be analyzed in POMAShiny. Data format must be a .CSV *comma-separated-value* file.

0.0.1.1 Target File

A .CSV with two mandatory columns (+ optional covariates):

- Each row denotes a sample (the same as in the features file)
- First/Left-hand column must be sample IDs => red
- Second/Left-hand column must be sample group/factor (e.g. treatment) => green
- Covariates (optional): From the third column (included) users can also include several experiment covariates => purple

Once this file has been uploaded, the user can select desired rows in the "Target File" panel table to create a subset of the whole uploaded data. If this selection is done, only selected rows will be analyzed in POMAShiny, if not (default) all uploaded data will be analyzed.

0.0.1.2 Features File

A .CSV with m columns:

- Each row denotes a sample and each column denotes a feature
- First row must contain the feature names

0.0.1.3 Exploratory report

After uploading the data and clicking the "Submit" button, POMAShiny allows users to generate a PDF exploratory data analysis report automatically by clicking the green button with the label "Exploratory report" in the top of the central panel. See an example here.

0.0.1.4 Example data

POMAShiny includes two example datasets that are both freely available at https://www.metabolomicsworkbench.org. The first example dataset consists of a targeted metabolomics three-group study and the second example dataset consists of a targeted metabolomics two-group study. These two datasets allow users to explore all available functionalities in POMAShiny. Both dataset documentations are available at https://github.com/pcastellanoescuder/POMA.

NOTE: Once target and features files are uploaded and the desired rows are selected in the target file (if necessary), users must have to click the "Submit" button to continue with the analysis.

Equivalent functions in POMA: `POMA::PomaMSnSetClass()` (format data) and `POMA::PomaEDA()` (automatic PDF report).

0.0.2 Pre-processing Panel

0.0.2.1 Impute Values

Usually, mass spectrometry faces with a high number of missing values, most of them due to low signal intensity of peaks. Missing value imputation process in POMAShiny is divided in three sequential steps:

1. Distinguish between zeros and missing values. In case the data have values of these two types users can distinguish or not between these values. This option may be useful in experiments combining endogenous and exogenous features, as in this case the exogenous ones could be a real zero and the endogenous ones are unlikely to be real zeros.
2. Remove all features of the data that have more of a specific percentage (defined by user) of missing values in ALL study groups. By default this percentage is 20%.
3. Imputation. POMAShiny offers six different methods to impute missing values:

- replace missing values by zero
- replace missing values by half of the minimum positive value in the original data (in each column)
- replace missing values by the median of the column (feature)
- replace missing values by the mean of the column (feature)
- replace missing values by the minimum value in the column (feature)
- replace missing values using KNN algorithm (default)

Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á., & Barbas, C. (2015). Missing value imputation strategies for metabolomics data. Electrophoresis, 36(24), 3050-3060.

Equivalent function in POMA: `POMA::PomaImpute()`.

0.0.2.2 Normalization

It's known that some factors can introduce variability in MS data. Even if the data have been generated under identical experimental conditions, this introduced variability can have a critical influence on the final statistical results, making normalization a key step in the workflow.

POMAShiny offers six different methods to normalize data:

- Autoscaling
- Level scaling
- Log scaling
- Log transformation
- Vast scaling
- Log pareto scaling (default)

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics, 7(1), 142.

Users can evaluate the normalization effects in the interactive boxplots located in the "Normalized Data" tab.

Equivalent functions in POMA: `POMA::PomaNorm()` (normalization) and `POMA::PomaBoxplots(group = "samples")` (boxplots).

0.0.2.3 Outlier Detection

POMAShiny allows the analysis of outliers by different plots and tables as well as the possibility to remove statistical outliers from the analysis (default) using different modulable parameters.

The method implemented in POMAShiny is based on the euclidean distances (default but modulable) among observations and their distances to each group centroid in a two-dimensional space. Once this is computed, the classical univariate outlier detection formula $Q3 + 1.5 * IQR$ (coefficient is modulable by the user) will be used to detect multivariate group-dependant outliers using computed distance to each group centroid.

Select the distance, type and coefficient to adapt the outlier detection method to your data. By switching the button "Show labels" all plots will display automatically the sample IDs in the outlier detection plots.

- Distances Polygon Plot: Group centroids and sample coordinates in a two-dimensionality space
- Distances Boxplot: Boxplots of all computed distances to group centroid by group

NOTE: If the "Remove outliers" button is turned on (default), all detected outliers are excluded from the analysis automatically.

Equivalent functions in POMA: `POMA::PomaOutliers(do = "analyze")` (to analyze outliers) and `POMA::PomaOutliers(do = "clean")` (to remove outliers).

0.0.3 EDA Panel

POMAShiny offers several interactive and highly modulable plots designed to facilitate the exploratory data analysis (EDA) process, giving a wide range of visualization options.

0.0.3.1 Volcano Plot

In this tab, users can explore their data in an interactive volcano plot. This plot is based on the results of a standard T-test gives information about This option is only available for two-group studies.

0.0.3.2 Boxplot

0.0.3.3 Density Plot

0.0.3.4 Heatmap

0.0.4 Statistical Analysis Panel

0.0.4.1 Univariate Analysis

Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships.

0.0.4.1.1 Parametric Tests

T-test

This is an statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. This analysis is used when you are comparing **two groups.**

A t-test is applied when the variable follow a **normal distribution**.

Correlated (or Paired) T Test: The correlated T test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures. This method can also applies on cases where the samples are related in some manner or have matching characteristics. Correlated or paired T tests are of dependent type, as these involve cases where the two sets of samples are related.
Equal Variance (or pooled) T Test: The equal variance T test is used when the number of samples in each groups is the same, or the variance of the two data sets is similar.

Unequal Variance T Test: The unequal variance T test is used when the number of samples in each group is different and the variance of the two data sets is also different. This test is also called the **Welch's t-test**.

ANOVA

A variance analysis (ANOVA) tests the hypothesis that the averages of **two or more** populations are the same. The ANOVA evaluates the importance of one or more factors when comparing the means of the response variable in the different levels of the factors. The null hypothesis states that all the means of the population (mean of the levels of the factors) are the same, while the alternative hypothesis states that at least one is different.

In this method you can analyze your **Covariates file** if you have it.

0.0.4.1.2 Non-parametric Tests

Mann-Whitney U Test

Mann-Whitney U test is the **non-parametric alternative test to the independent sample t-test**. It is a non-parametric test that is used to compare **two group** means that come from the same population, and used to test whether two sample means are equal or not. Usually, the Mann-Whitney U test is used when the data is ordinal or when the **assumptions of the t-test are not met**.

When your have **paired groups**, this test becomes a **Wilcoxon signed-rank test**.

Assumptions:

1. The sample drawn from the population is random.
2. Independence within the samples and mutual independence is assumed. That means that an observation is in one group or the other (it cannot be in both).
3. Ordinal measurement scale is assumed.

Kruskal Wallis Test

Kruskal-Wallis test is a non-parametric method to test whether a group of data comes from the same population. Intuitively, it is similar to the ANOVA with the data replaced by categories. It is an extension of the Mann-Whitney U test for **3 or more groups**.

Since it is a non-parametric test, the Kruskal-Wallis test does not assume normality in the data, as **opposed to the traditional ANOVA**. It assumes, under the null hypothesis, that the data come from the same distribution.

0.0.4.2 Multivariate Analysis

0.0.4.2.1 PCA

0.0.4.2.2 PLS-DA

0.0.4.2.3 sPLS-DA

0.0.4.3 Cluster Analysis

0.0.4.3.1 k-means

0.0.4.3.2 MDS

0.0.4.4 Limma

Limma (Linear Models for Microarray Data) is an R/Bioconductor software package that provides an integrated solution for analysing high-throughput experimental data. It contains rich features for information borrowing to overcome the problem of **small sample sizes**.

How it works?

On one hand, it fits a linear model to each feature of data and takes advantage of the flexibility of such models in various ways, for example to handle complex experimental designs and to test very flexible hypotheses.

On the other hand, it leverages the highly parallel nature of features to borrow strength between the feature-wise models, allowing for different levels of variability between features and between samples, and making statistical conclusions more reliable when the number of samples is small.

In this method you can analyze your **Covariates file** if you have it.

You have to normalize the data to use this method.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.

0.0.4.5 Correlation Analysis

0.0.4.6 Lasso

0.0.4.6.1 Lasso

0.0.4.6.2 Ridge Regression

0.0.4.6.3 Elasticnet

0.0.4.7 Random Forest

0.0.4.8 Rank Products

0.0.4.9 Odds Ratio