Replication of Partial Least Squares Discriminant Analysis on Metabolomics NMR/MS Data

Florencia Marcaccio
Sanjana Gupta
Medha Sagar
Adrian Tullock
Faryal Usman

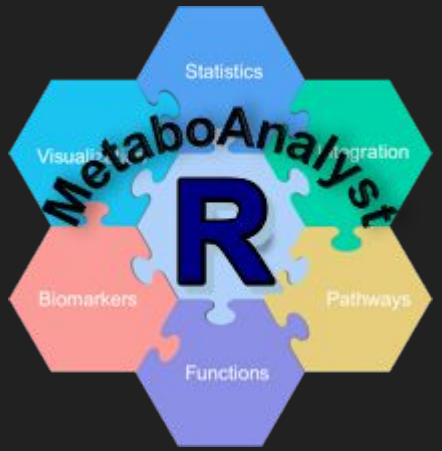
Publication Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703878/

Repository:
https://github.com/UW-MSDS-DATA-598-Reproducibility-WI20/gupta-marcaccio-sagar-tullock-usman-replication-project

Introduction

- The purpose of our paper is to serve as the home of the replication project for the metabolomics analysis
- Metabolomics is the scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates and products of metabolism.
- There are a number of data processing challenges germane to metabolomics
- These processes include metabolomic data processing, normalization, multivariate statistical analysis, and data annotation

Target Paper - MetaboAnalyst: a web server for matabolomic data analysis and interpretation



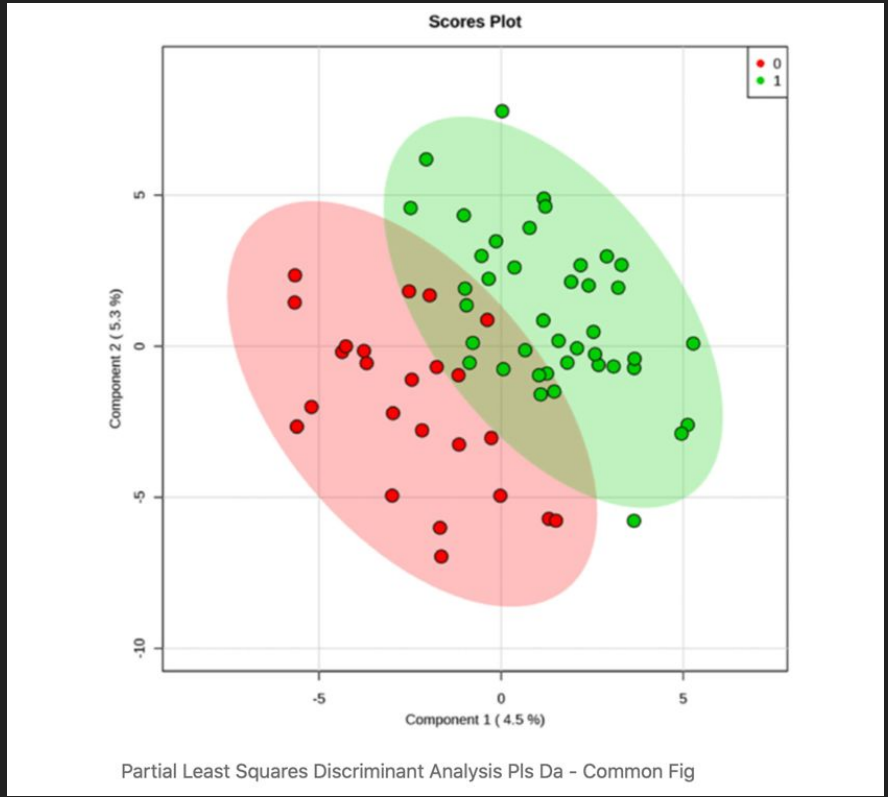
- Web-based tool for metabolomics studies.
- Accepts a variety of input data.
- Four modules:
  - Data processing
  - Statistical analysis
  - Functional enrichment analysis
  - Metabolic pathway analysis

Partial Least Square Discriminant Analysis

Data used in the Metabolomics field:

- Highly correlated
- High-dimensional
- Few observations
- No knowledge of underlying distribution

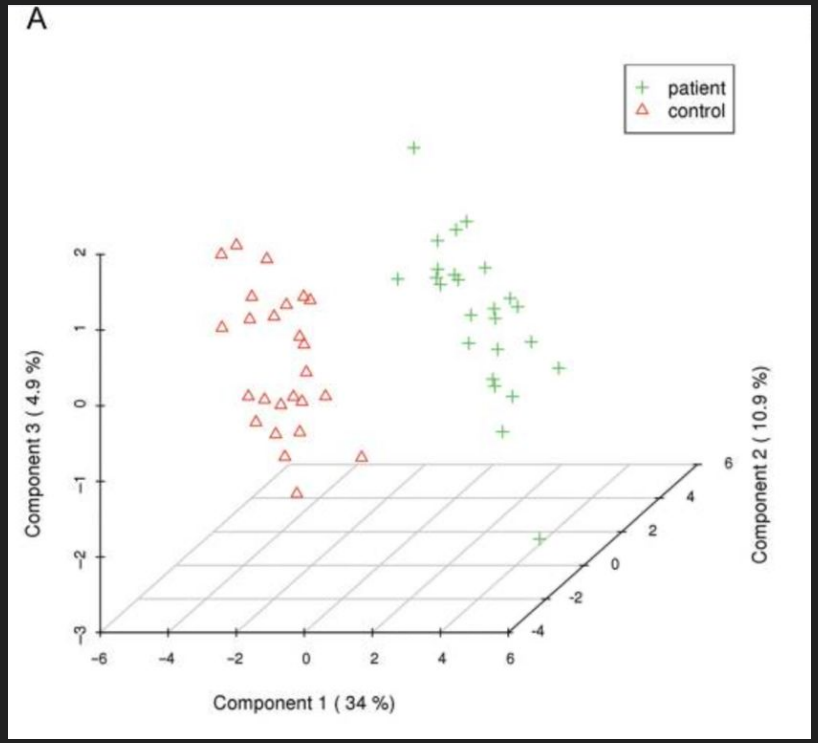
Traditional linear methods fail, so **PLS-DA** and **PCA** are used instead



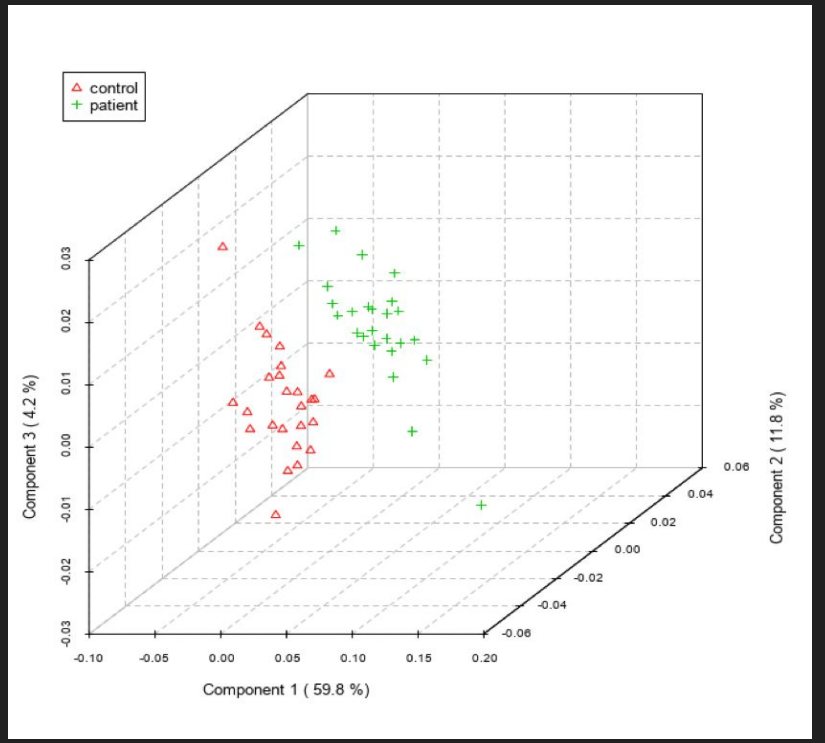
Partial Least Squares Discriminant Analysis Pls Da - Common Fig

Partial Least Square Discriminant Analysis

Target figure



Obtained figure



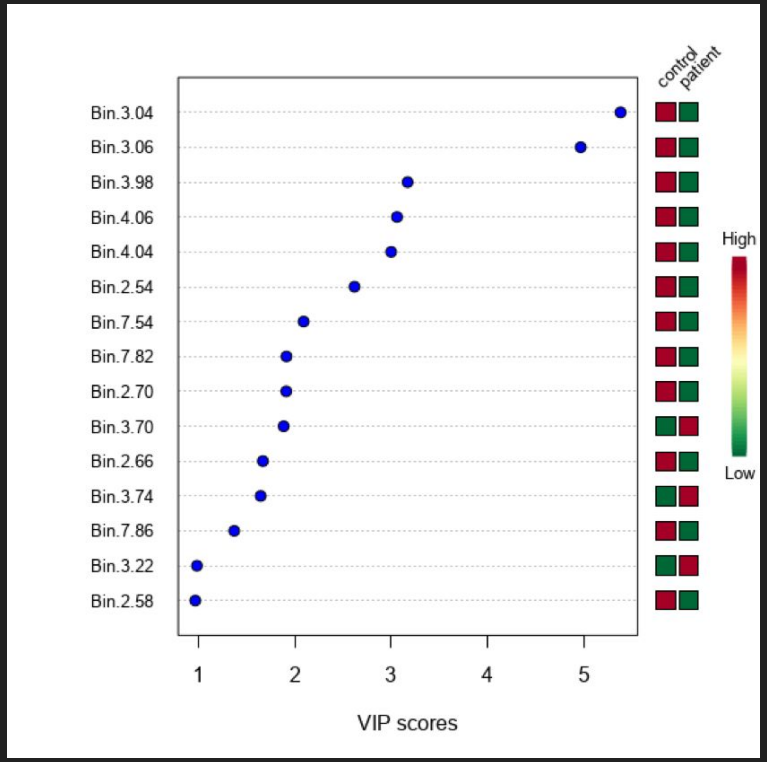
Technical Variations

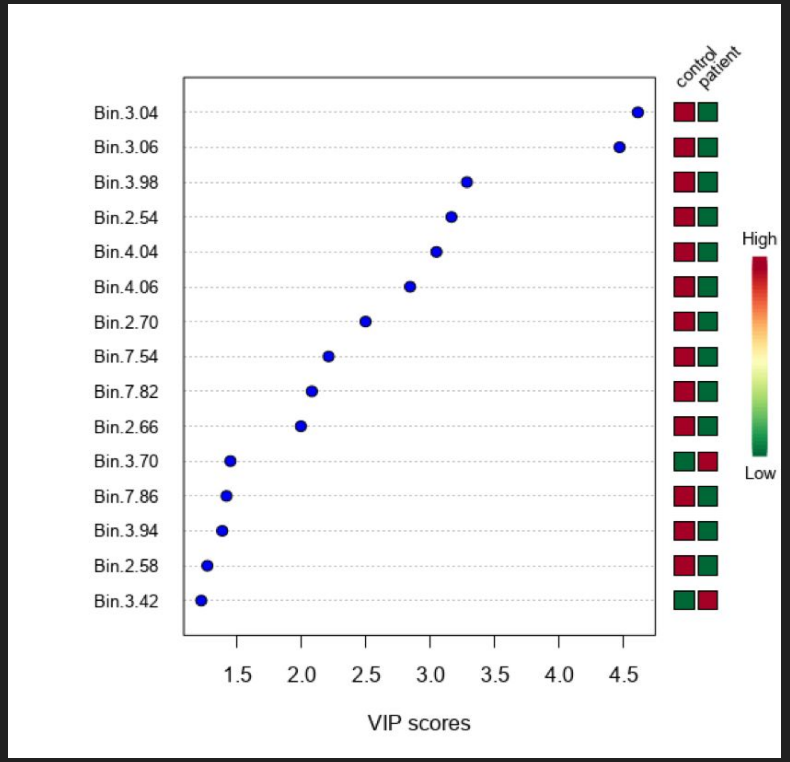
- Imputation method of missing values
- Filtering method of unused variables

} Did not make a difference due to already clean data

- **Component for variable importance**

Technical Variations: Component for variable importance





Challenges

Domain Knowledge

Not many packages produce reliable PLSDA score plots.

We unsuccessfully tried the following packages:

- DiscriMiner
- GGPlot2 PCA
- PLSDA
- mdatools

Technical

- Problems installing MetaboAnalyst package.
- Re-defining the native data types to the package is not possible.

- Install CAMERA through BioConductor manually
- Install XQuartz for compatibility with the Cairo package

Theoretical

Not enough information on data processing methods used in the original paper.

While we tried different combinations, our replicated figure is similar but not exactly the same as target figure.

Conclusions

- We experienced some difficulties when installing the package.
- The processing steps are not detailed enough in the original paper.
- While we were able to replicate the figure in the target paper, there are noticeable changes between both figures. We believe that this is due to some outlier removal technique not mentioned in the paper.