Cell-wise outlier diagnostics based on pairwise log-ratios

Jan Walach[1],P.Filzmoser[1], K.Hron[2], Š.Kouřil[3,4]

[1]TU Wien, [2]Palacky University Olomouc, [3]University Hospital Olomouc

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

June 26, 2018

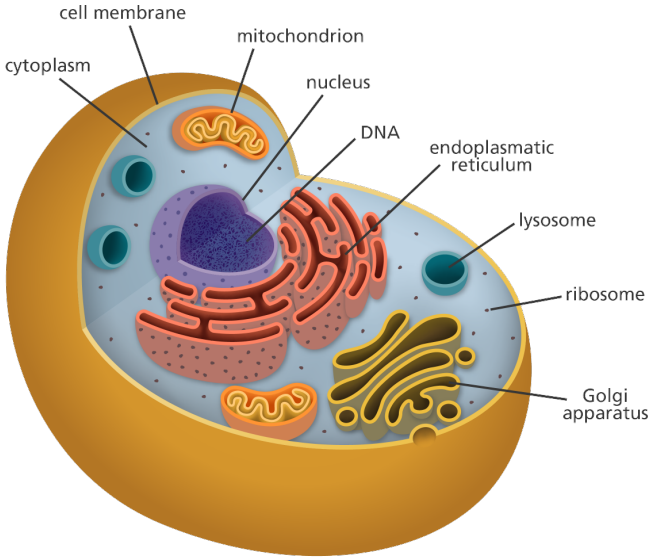
Data

- Focus on metabolomics data
- Two groups (e.g. control/disease patients)
- Usually low number of observations
- Flat data: many variables

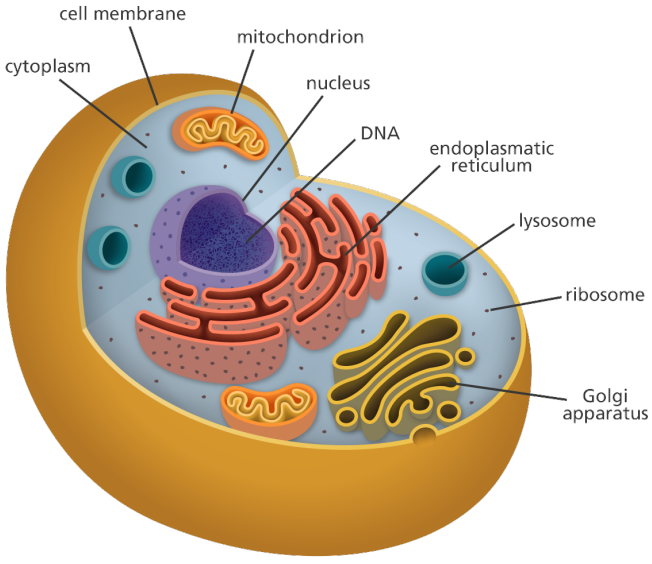
Goal

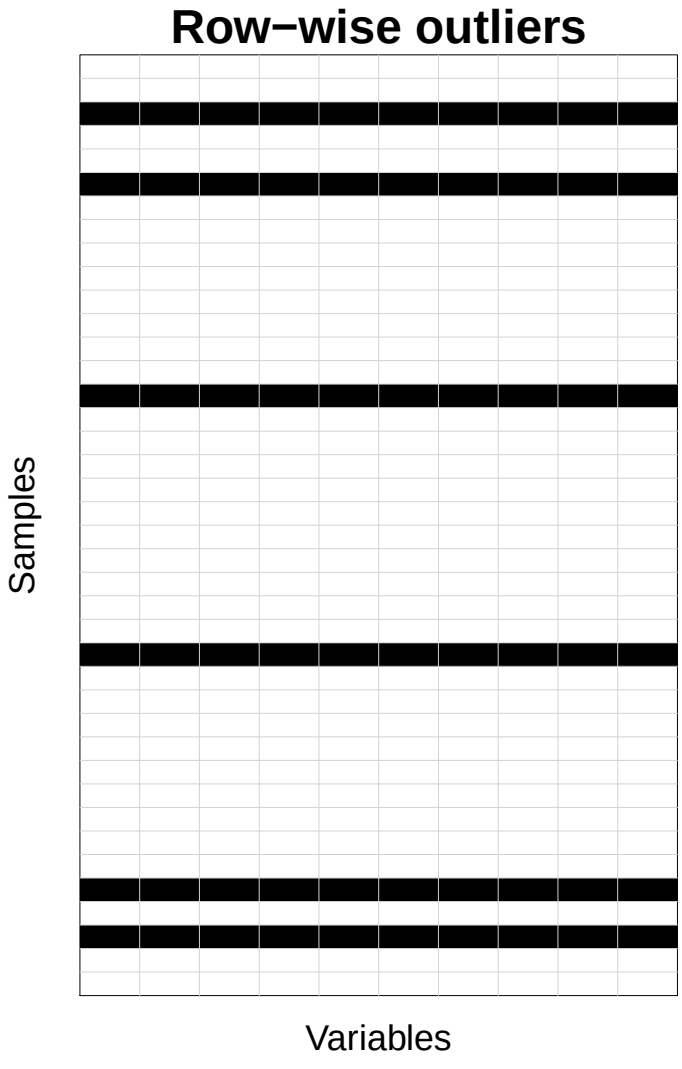
- Clustering/Classification
- Exploratory analysis
- Identification of non-standard samples/values
- Find informative variables

Biomarker identification

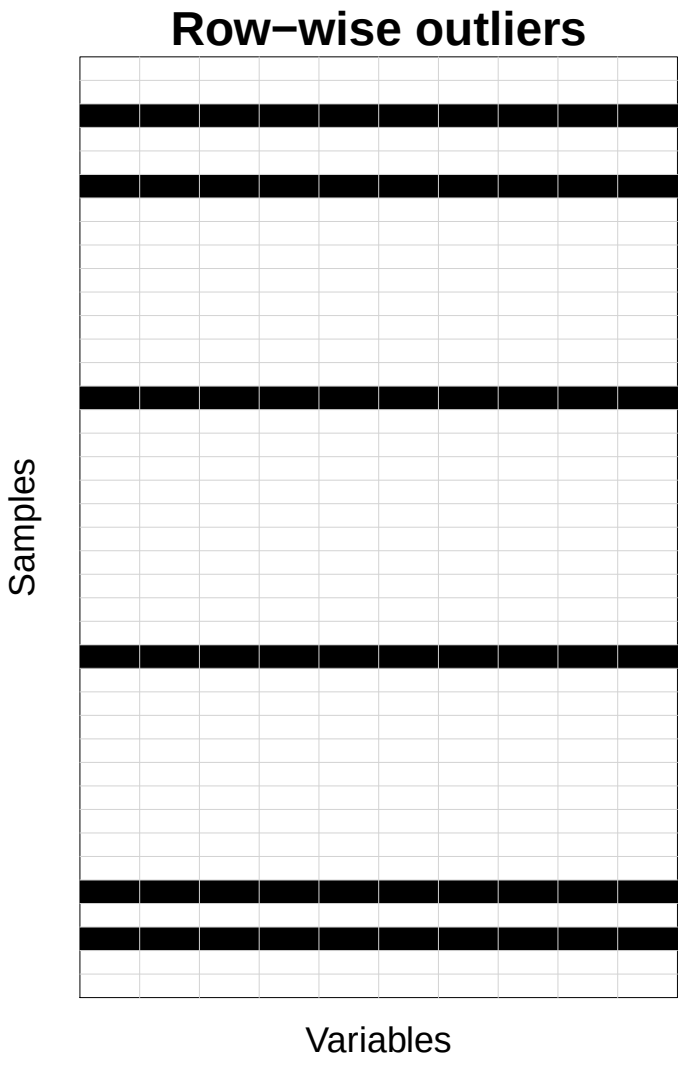


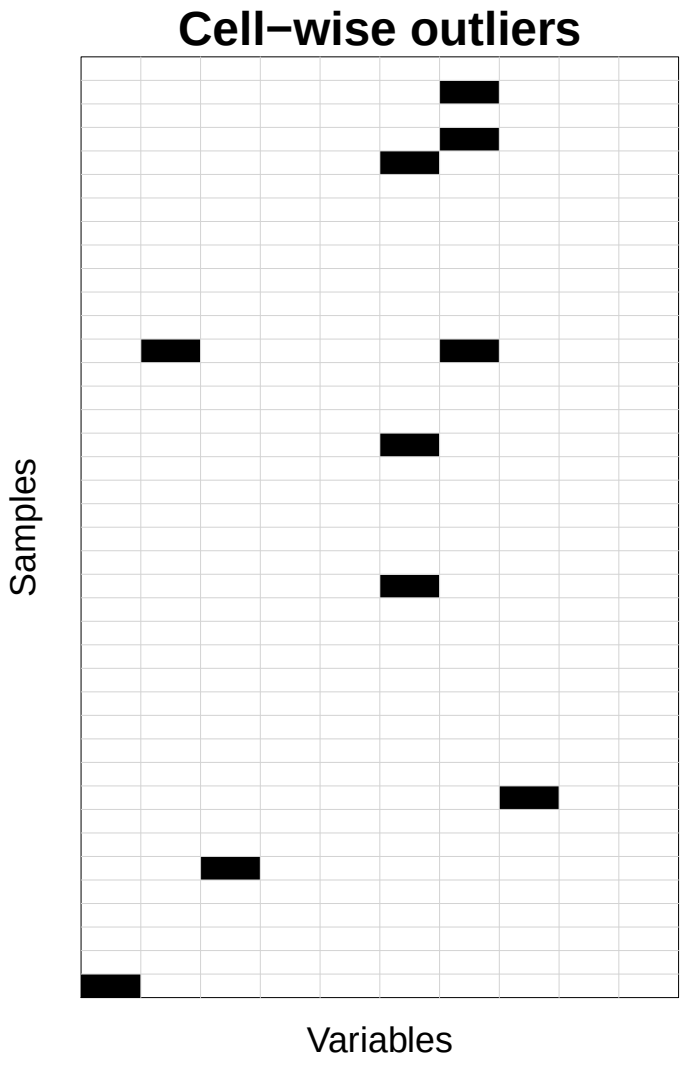
Cell-wise outliers



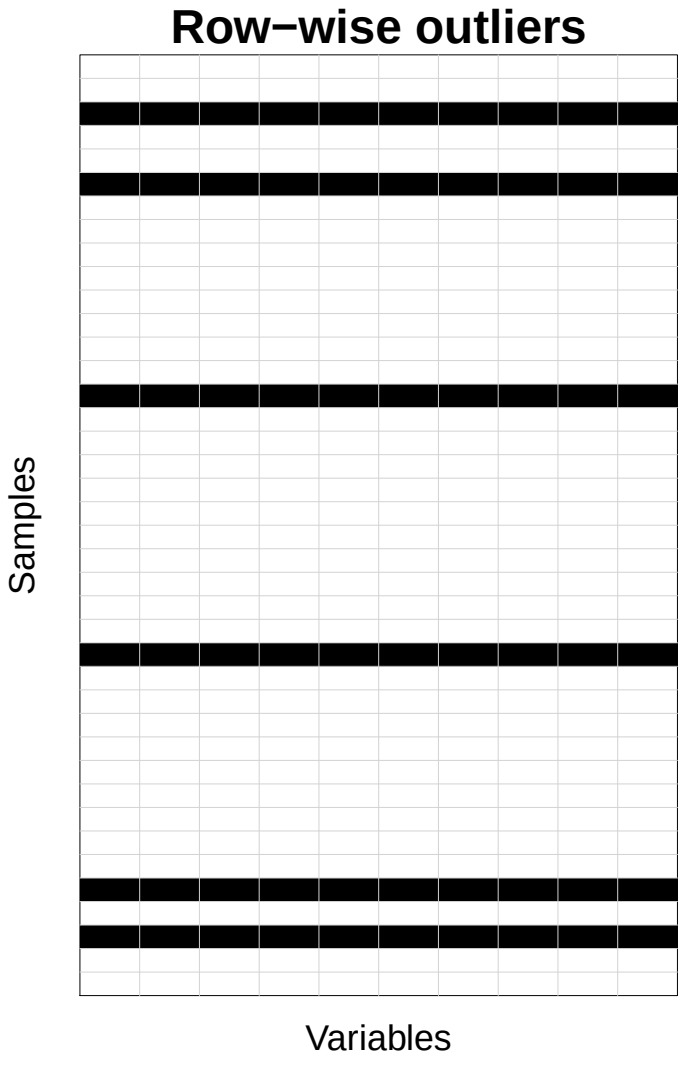

Row−wise outliers
Samples
Variables

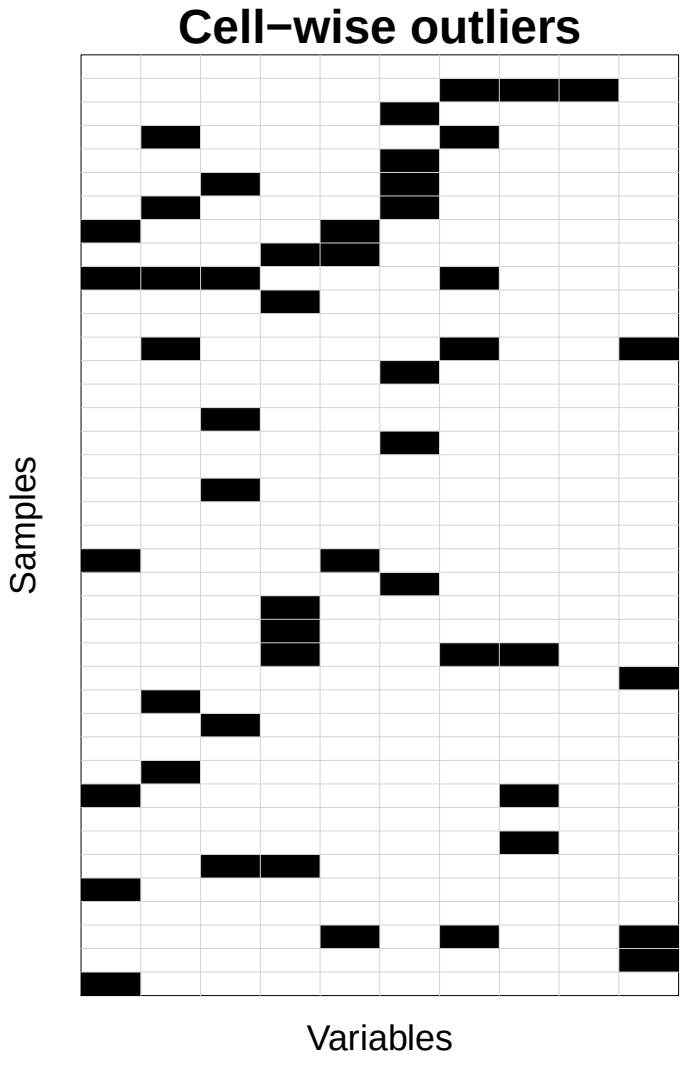
Outlier diagnostics





Outlier diagnostics





Cell-wise rPLR (1)

Steps:

1. Computing log-ratios
2. Robust centering + scaling
3. Applying weighting function
4. Projection to original space

Data

- Data matrix X with dimensions $n \times D$

Based on idea of robust variation matrix

- Matrix T, elements t_{jk}, for $j, k = 1, \ldots, D$

$$t_{jk} = \sigma \left[\ln \left(\frac{x_{1j}}{x_{1k}} \right), \ln \left(\frac{x_{2j}}{x_{2k}} \right), \ldots, \ln \left(\frac{x_{nj}}{x_{nk}} \right) \right], \qquad (1)$$

- where σ is estimation of scale
- There are in total $\frac{d \times (d-1)}{2}$ possible log-ratios
- Each ln() will be robustly centred + scaled and will get a weight

$$t_{jk} = \sigma \left[\ln \left(\frac{x_{1j}}{x_{1k}} \right), \ln \left(\frac{x_{2j}}{x_{2k}} \right), \ldots, \ln \left(\frac{x_{nj}}{x_{nk}} \right) \right],$$

Three dimentional weight matrix

- All weights can be stored in the three dimensional matrices $W^{(l)}$, with D rows, D columns and n_l slices
- Information about observations is in certain slices
- Information about cells in rows of slices

Diagnostics

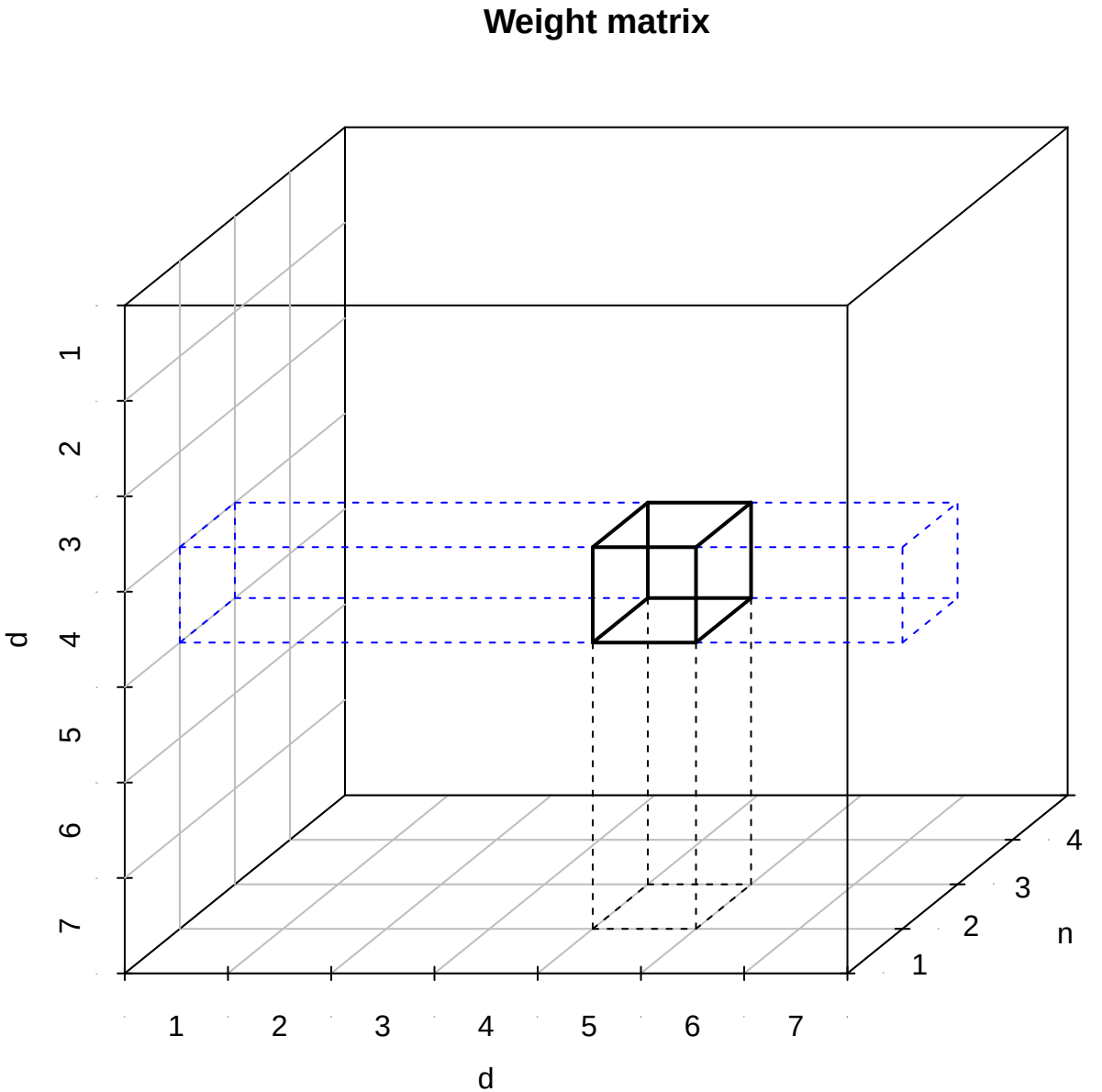
- Average of all weights for each observation and each involved variable,

$$m_{ij} = \frac{1}{D} \sum_{k=1}^{D} w_{jki} \tag{2}$$

- Or Median

$$m_{ij} = \underset{k}{\text{median}} \left(w_{jki} \right) \tag{3}$$

Projection to original dimensions

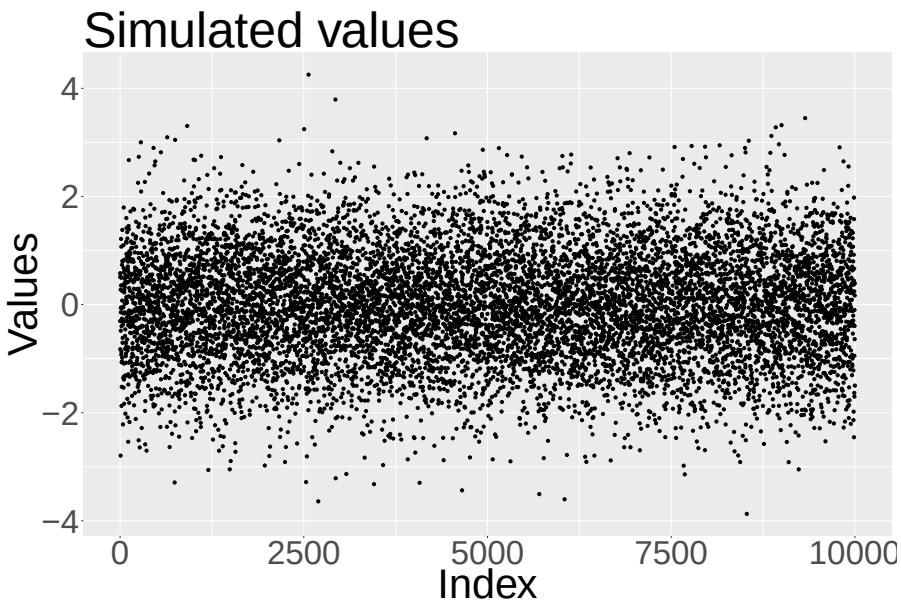


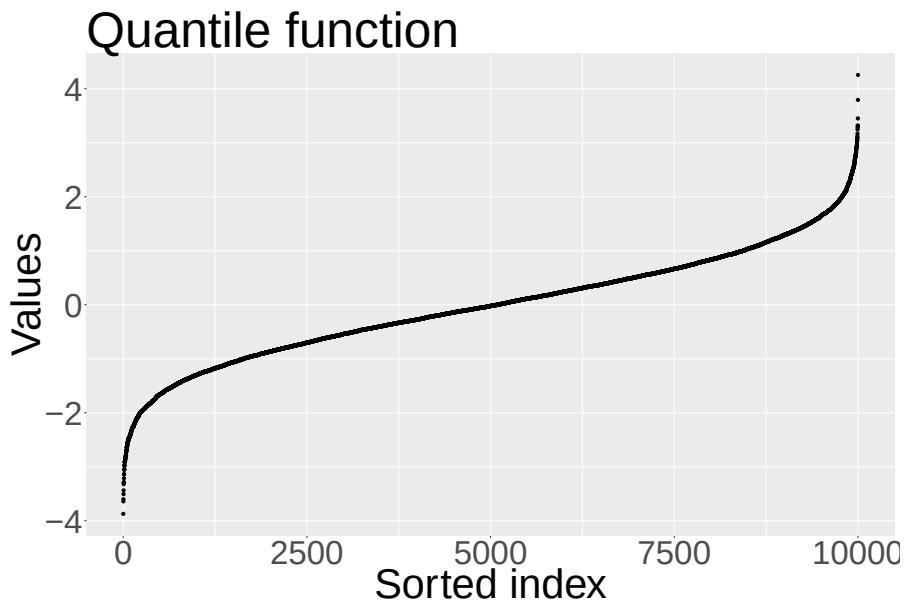
Applying weighting function

Weighting functions:

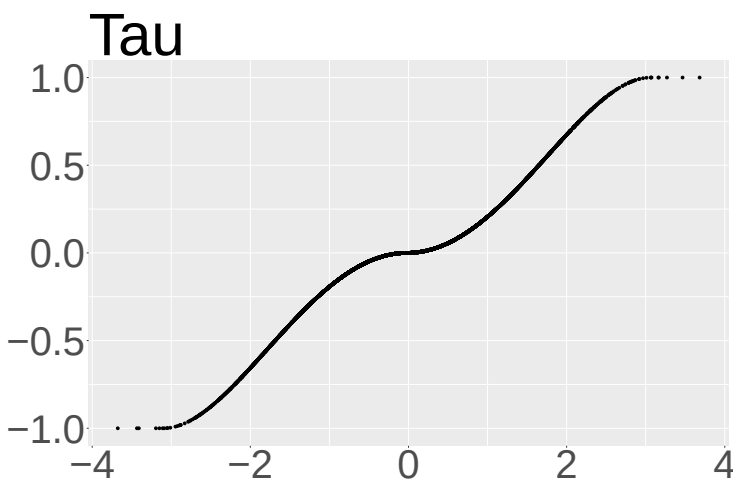
- τ function
- Tukey's biweight function
- Huber function
- Hampel function

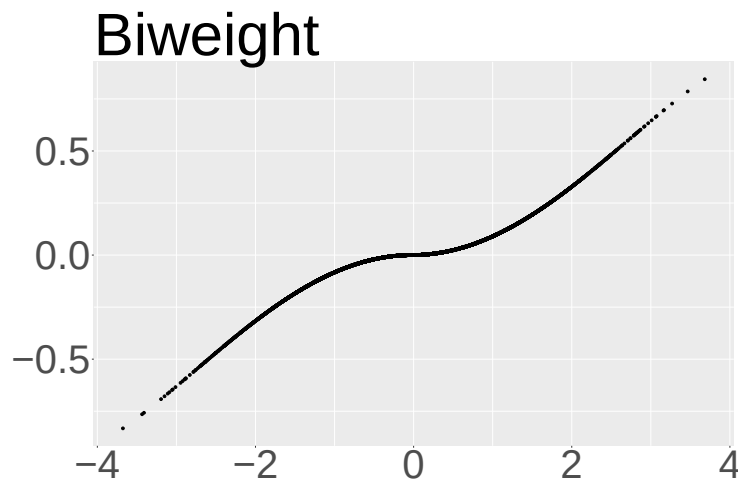
Outlier diagnostics

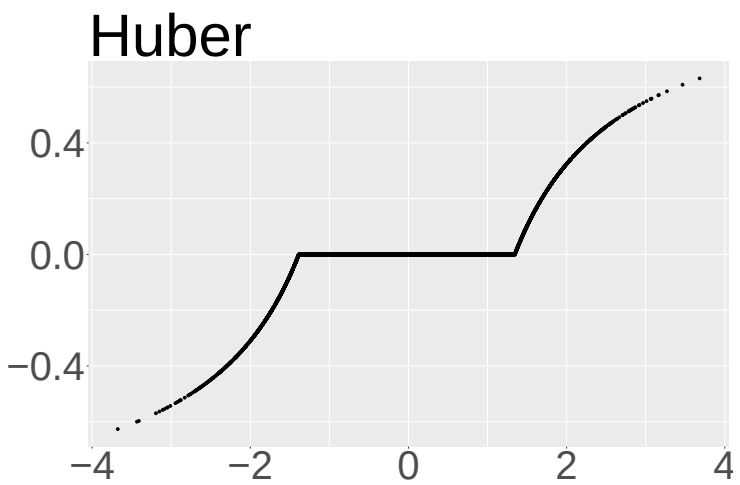


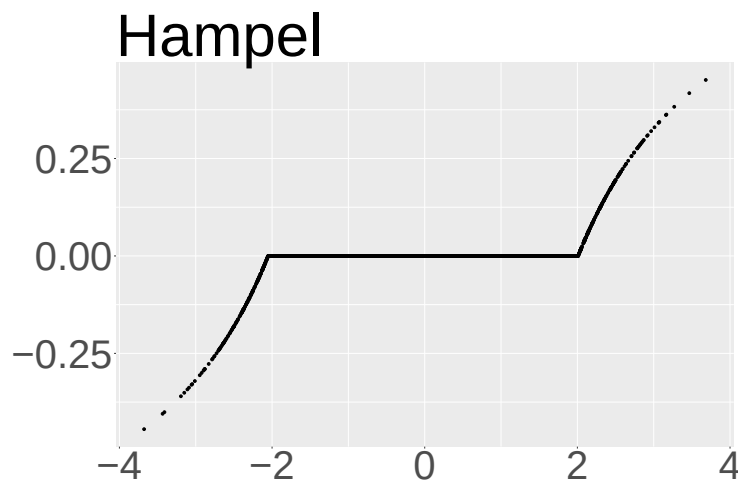


Cell-wise outliers



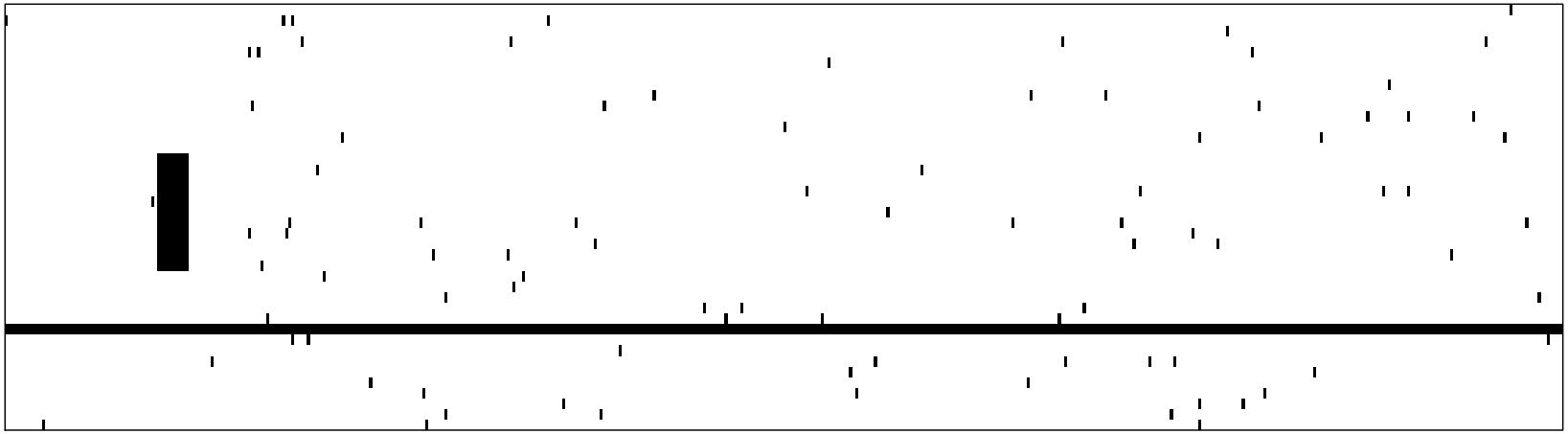




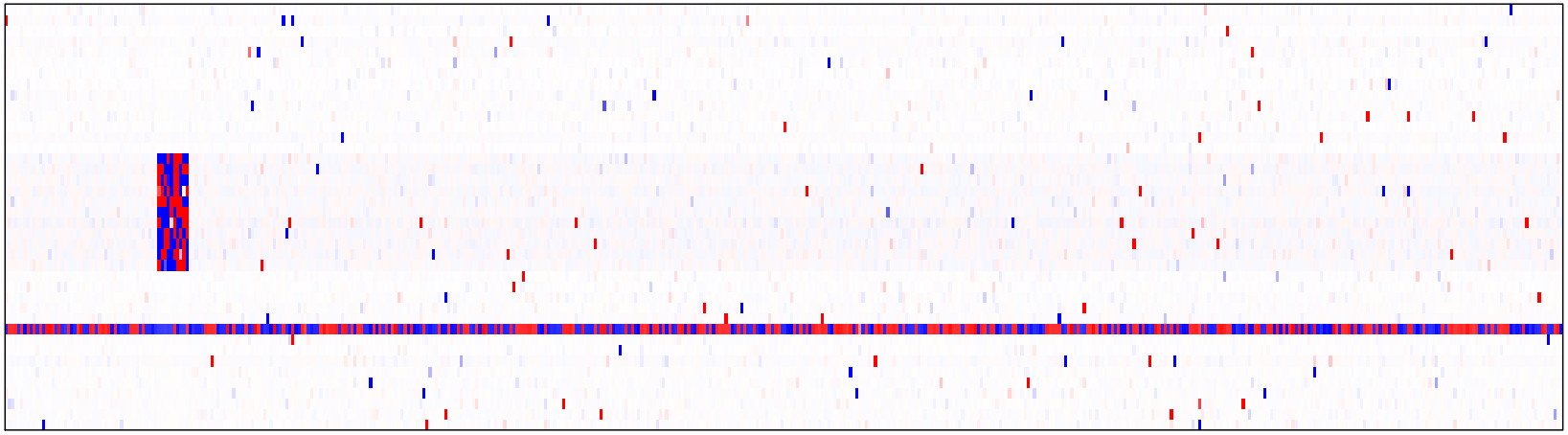


Diagnostics example: Mean

Imputed outliers

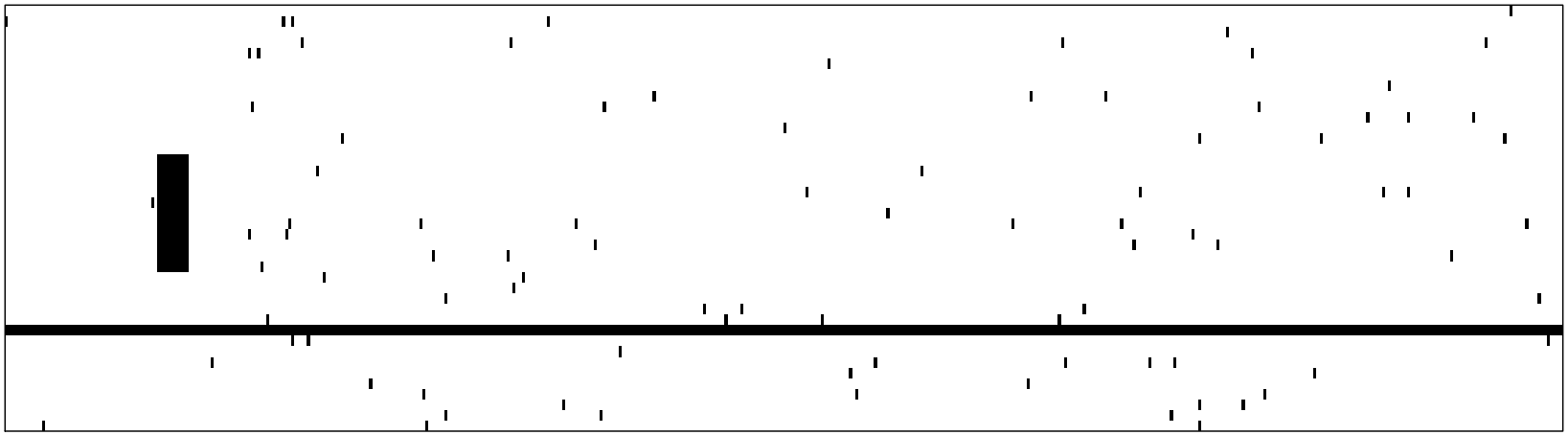


Identified outliers

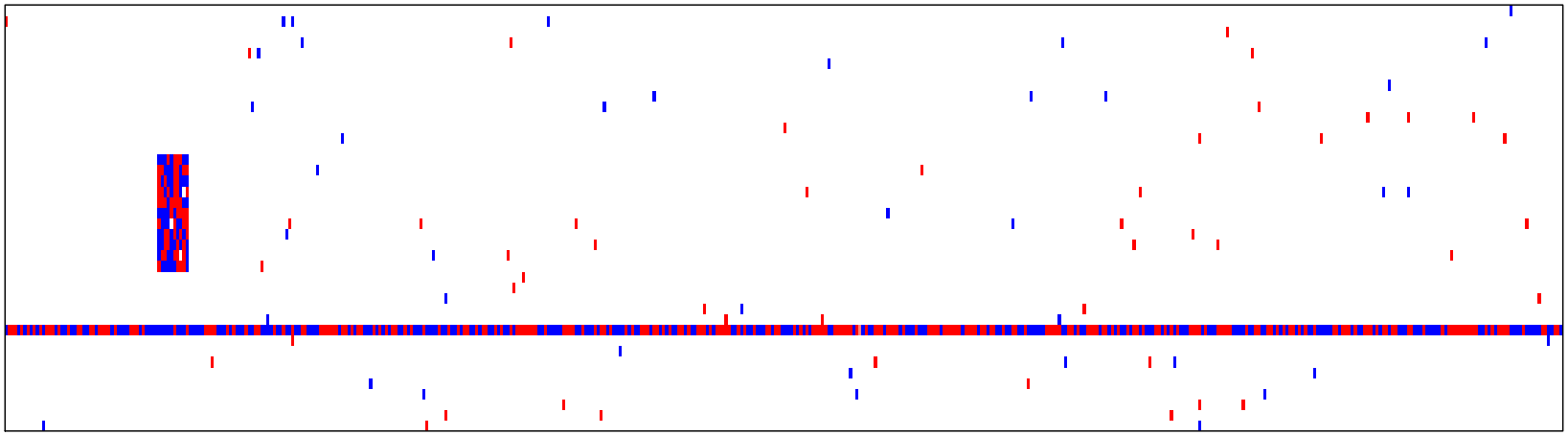


Diagnostics example: Median

Imputed outliers

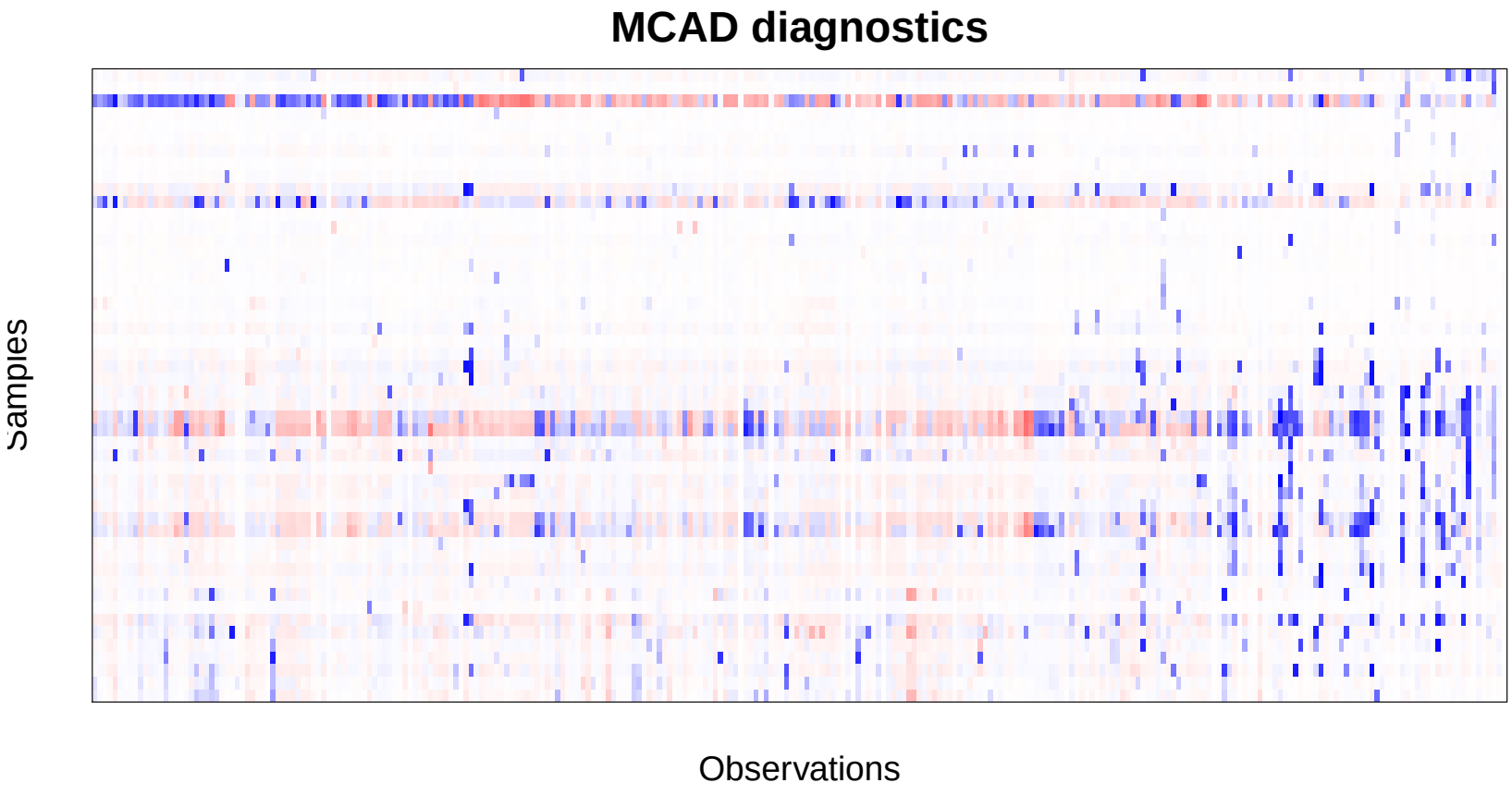


Identified outliers

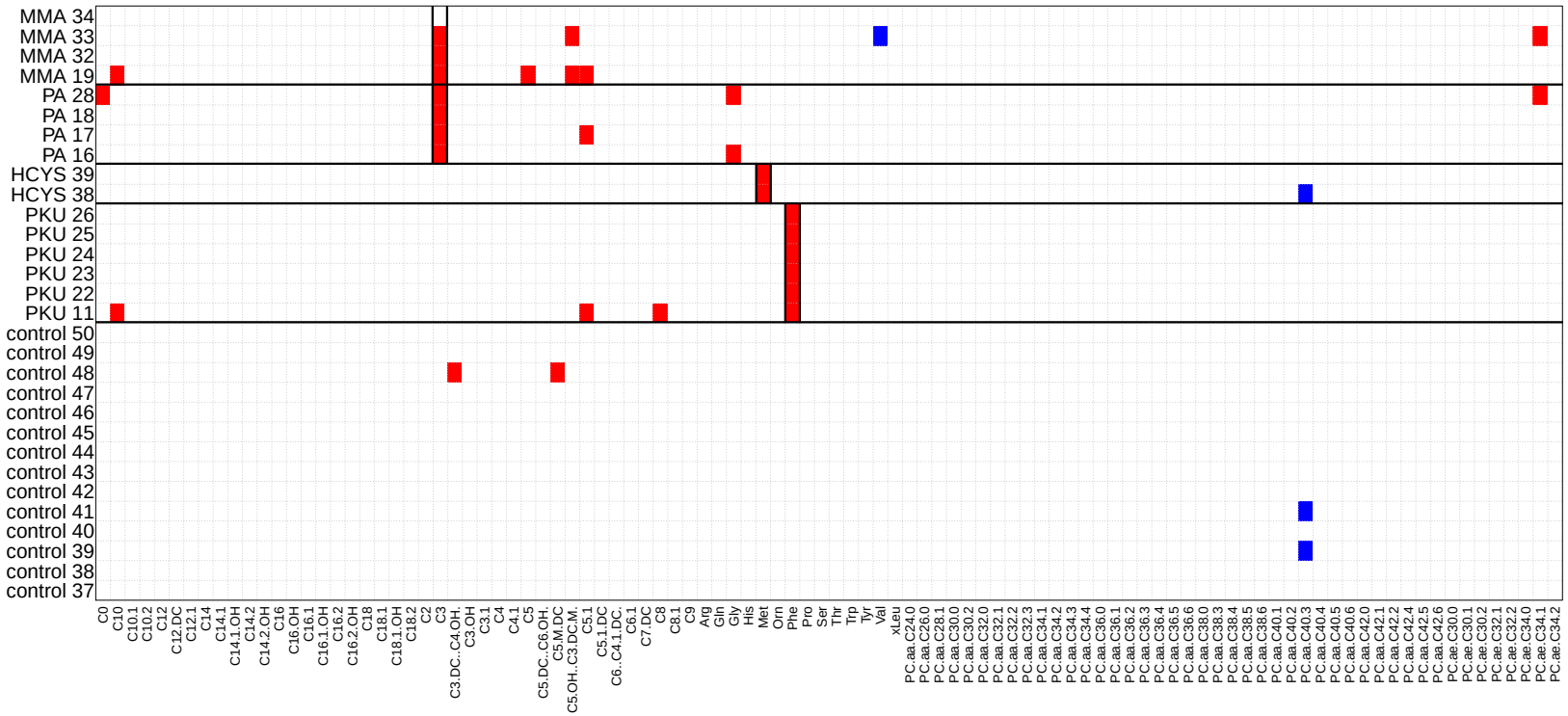


Real dataset MCADD

- MCADD (Medium chain acyl-CoA dehydrogenase deficiency)
- $n_1 = 25,\ n_2 = 25,\ D = 273$



Biomarkers = Cell-wise outliers in one variable for one group.



R-implementation

Cell-wise outliers

Threashold:

0 0.45 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Weighting function:

Hampel

Aggregation:

Median

Sorting:

By Number bigger than zero

Sort according to:

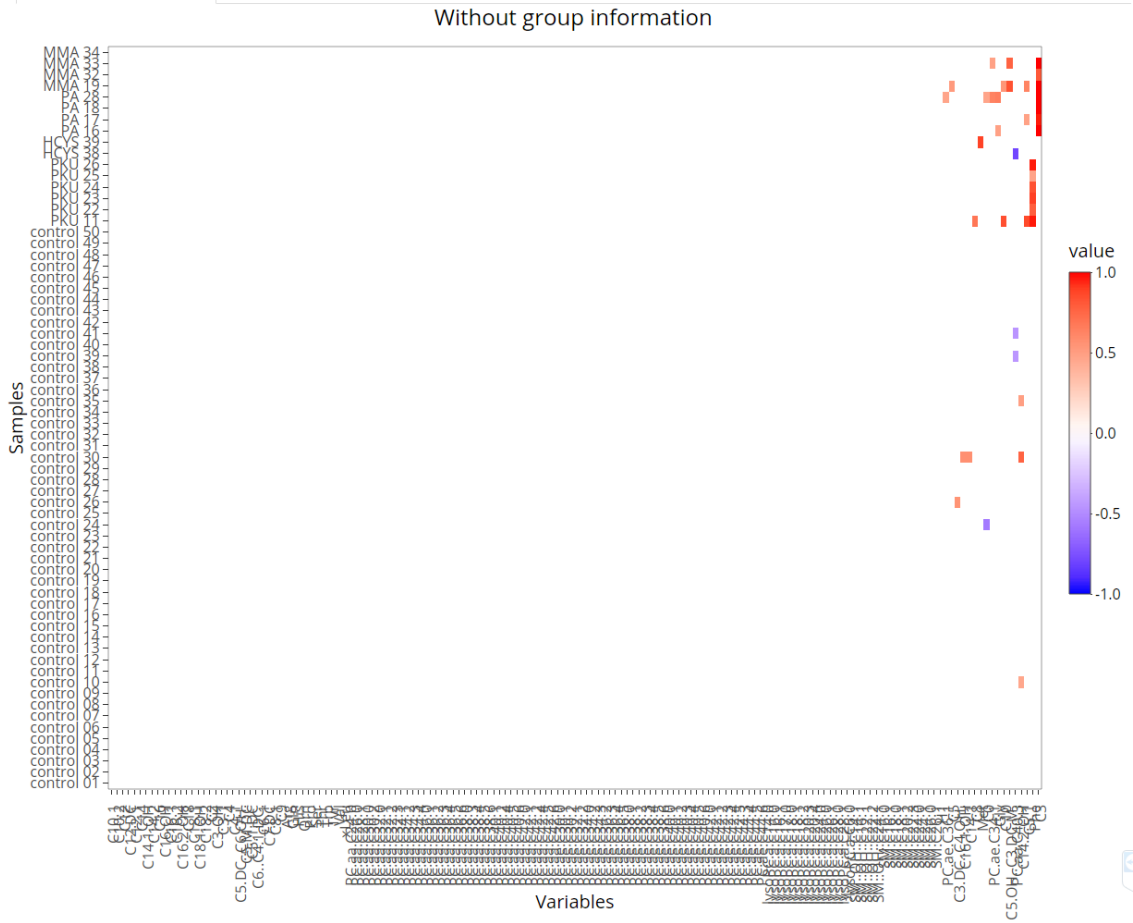
1 66

1 8 15 22 29 36 43 50 57 64 66

Filename

Download Plot

Without group information | With group information



Summary

- Introduction cell-wise outlier identification
- Possible use as a biomarker identification method

Limitations

- Computationally demanding in higher dimensions

Advantages

- Robust method
- Works also for highly unbalanced groups
- Easy to implement and apply:
  - `https://github.com/walachja/Cell-Wise_rPLR`

References

- Pawlowsky-Glahn, V., Egozcue, R. Tolosana-Delgado, J.J.: *Modeling and Analysis of Compositional Data.* Chichester: Wiley, 2015.
- Peter J. Rousseeuw, Wannes Van Den Bossche. "Detecting deviating data cells." Technometrics 60.2 (2018): 135-145.
- Walach, J., Filzmoser, P., Hron, K., Walczak, B., Najdekr, L.: Robust biomarker identification in a two-class problem based on pairwise log-ratios. *Chemometrics and Intelligent Laboratory Systems*, 171 (2017), 277-285
- P. Filzmoser, B. Walczak, What can go wrong at the data normalization step for identification of biomarkers?, Journal of Chromatography A 1362 (2014) 194–205.
- J. Walach, P. Filzmoser, K. Hron, Š. Kouřil ,Cell-wise outlier diagnostics based on log-ratio methodology, In preparation
- `https://github.com/walachja/Cell-wise_rPLR`

Real dataset