The use of log-ratio methodology in cell-wise outlier diagnostics

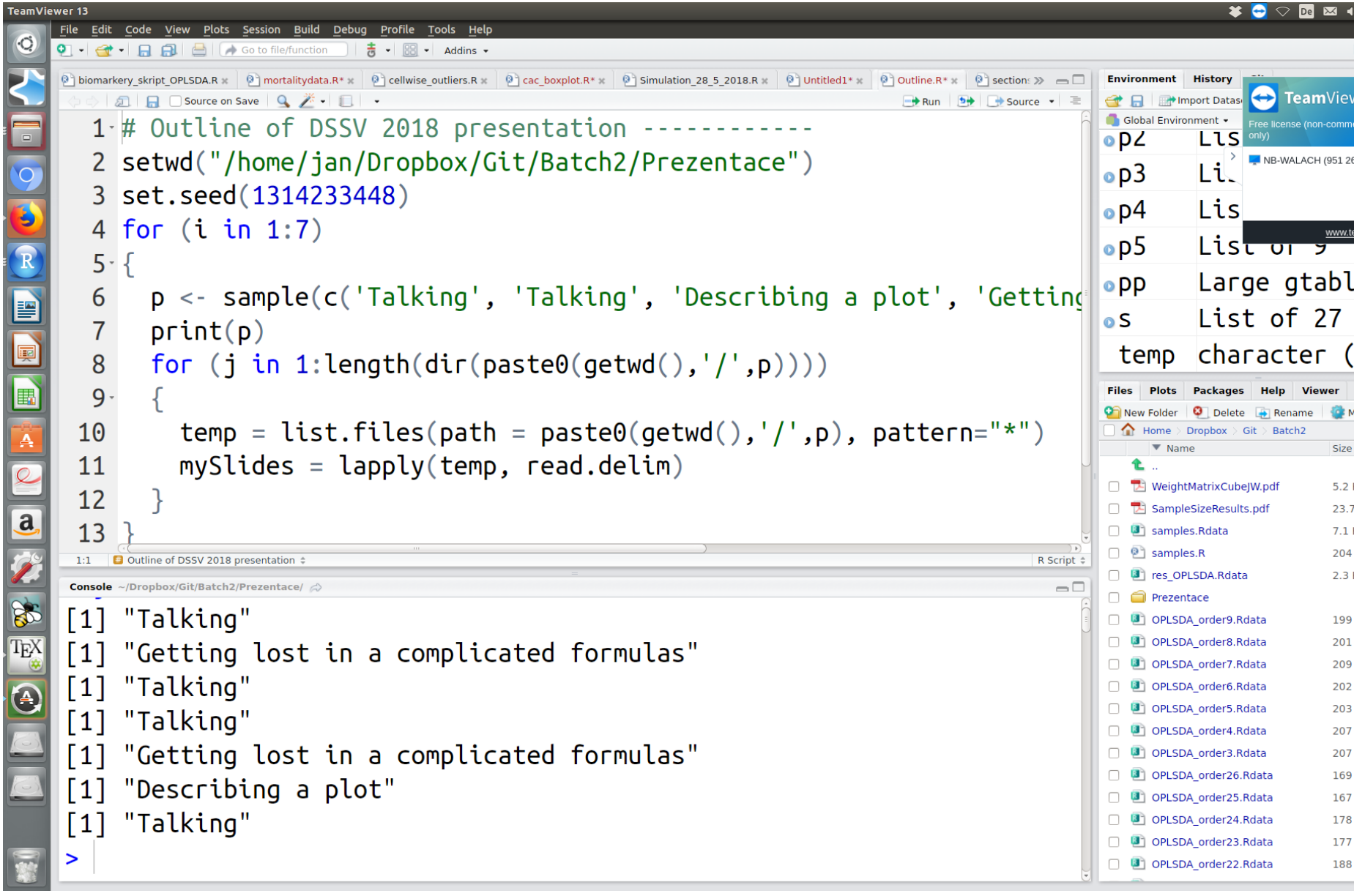
Jan Walach[1], P.Filzmoser[1], K.Hron[2]

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

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

June 11, 2018

Outline


TeamViewer 13
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
biomarkery_skript_OPLSDA.R
mortalitydata.R*
cellwise_outliers.R
cac_boxplot.R*
Simulation_28_5_2018.R
Untitled1*
Outline.R*
Source on Save
Run
Source
1 # Outline of DSSV 2018 presentation ------------
2 setwd("/home/jan/Dropbox/Git/Batch2/Prezentace")
3 set.seed(1314233448)
4 for (i in 1:7)
5 {
6 p <- sample(c('Talking', 'Talking', 'Describing a plot', 'Getting
7 print(p)
8 for (j in 1:length(dir(paste0(getwd(),'/',p))))
9 {
10 temp = list.files(path = paste0(getwd(),'/',p), pattern="*")
11 mySlides = lapply(temp, read.delim)
12 }
13 }
1:1
Outline of DSSV 2018 presentation
R Script
Console ~/Dropbox/Git/Batch2/Prezentace/
[1] "Talking"
[1] "Getting lost in a complicated formulas"
[1] "Talking"
[1] "Talking"
[1] "Getting lost in a complicated formulas"
[1] "Describing a plot"
[1] "Talking"
>
Environment History
Import Datas
Global Environment
p2 Lis
p3 Li
p4 Lis
p5 List of 9
pp Large gtabl
s List of 27
temp character (
Files Plots Packages Help Viewer
New Folder Delete Rename
Home Dropbox Git Batch2
Name Size
WeightMatrixCubeJW.pdf 5.2
SampleSizeResults.pdf 23.7
samples.Rdata 7.1
samples.R 204
res_OPLSDA.Rdata 2.3
Prezentace
OPLSDA_order9.Rdata 199
OPLSDA_order8.Rdata 201
OPLSDA_order7.Rdata 209
OPLSDA_order6.Rdata 202
OPLSDA_order5.Rdata 203
OPLSDA_order4.Rdata 207
OPLSDA_order3.Rdata 207
OPLSDA_order26.Rdata 169
OPLSDA_order25.Rdata 167
OPLSDA_order24.Rdata 178
OPLSDA_order23.Rdata 177
OPLSDA_order22.Rdata 188
TeamView
Free license (non-comm
only)
NB-WALACH (951 26

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

The use of log-ratio methodology in cell-wise outlier diagnostics

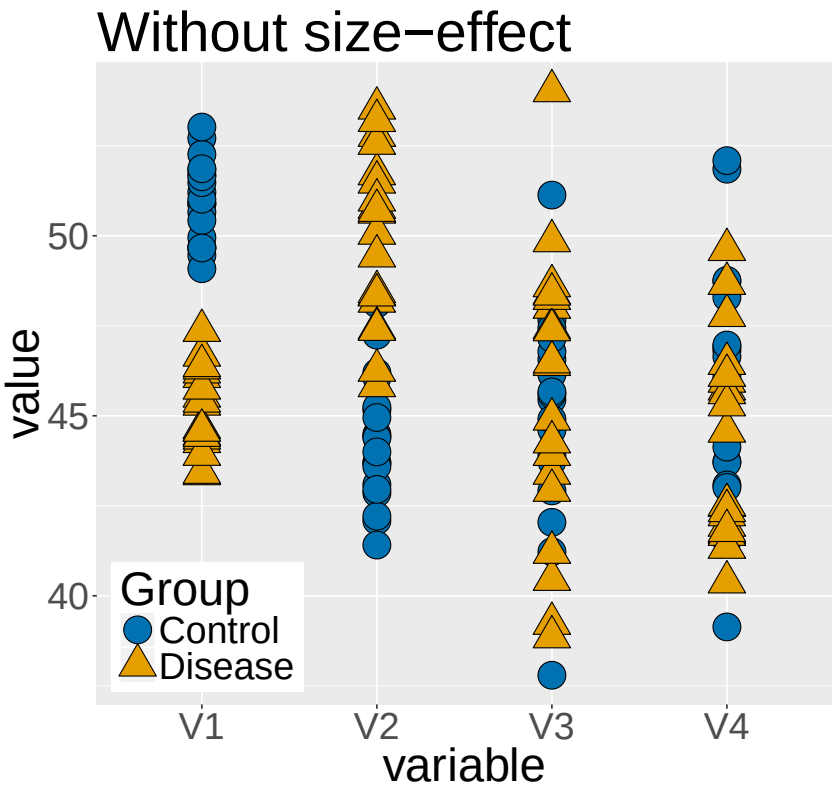
The use of log-ratio methodology in cell-wise outlier diagnostics

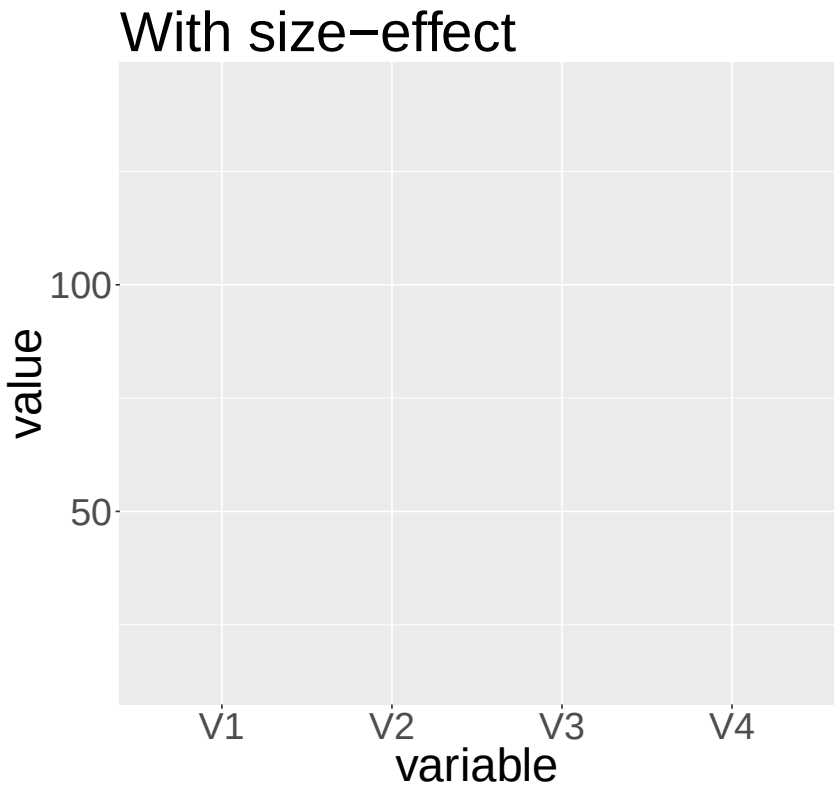
Pawlowsky-Glahn, V., Egozcue, R. Tolosana-Delgado, J.J.: *Modeling and Analysis of Compositional Data.* Chichester: Wiley, 2015.

Size-effect

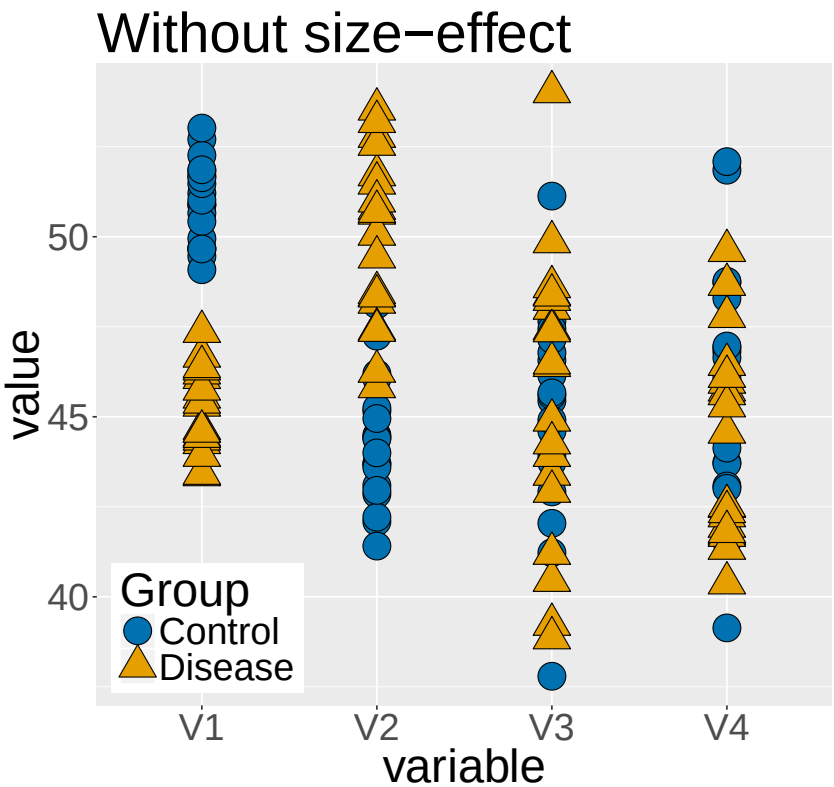
- Problem: "Size effect"
- Different sample volume and/or sample concentration
- Special treatment required

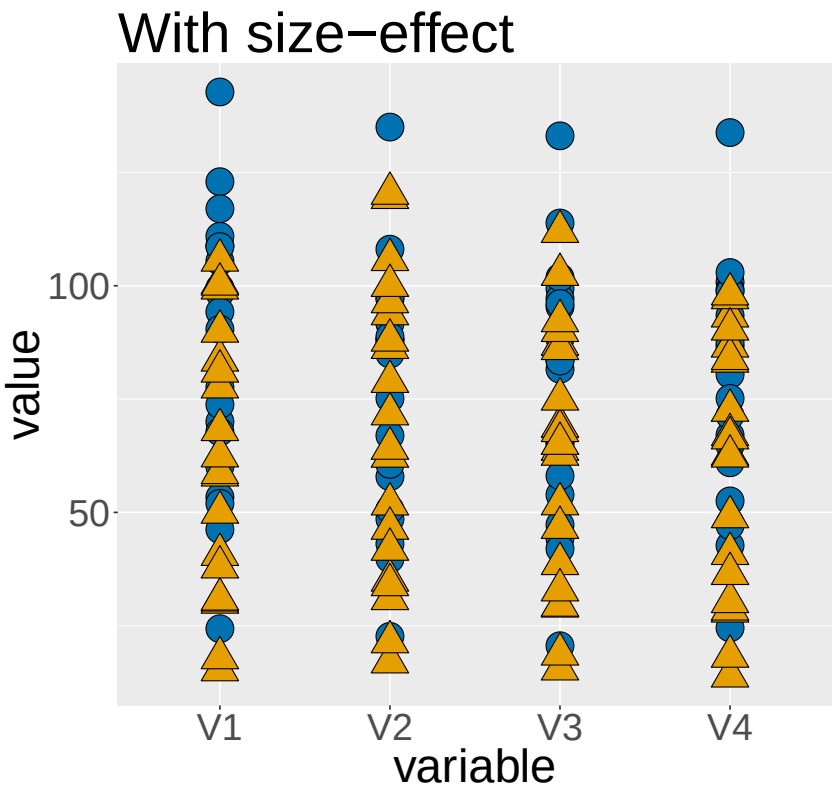
Size-effect


Without size−effect
value
50
45
40
Group
Control
Disease
V1
V2
V3
V4
variable

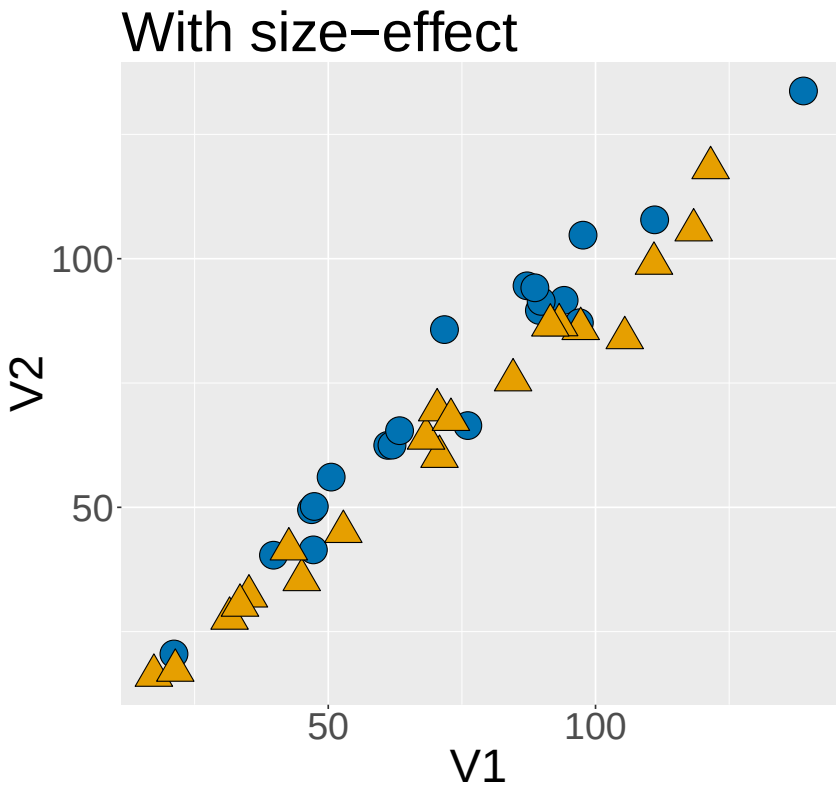

With size−effect
value
100
50
V1
V2
V3
V4
variable

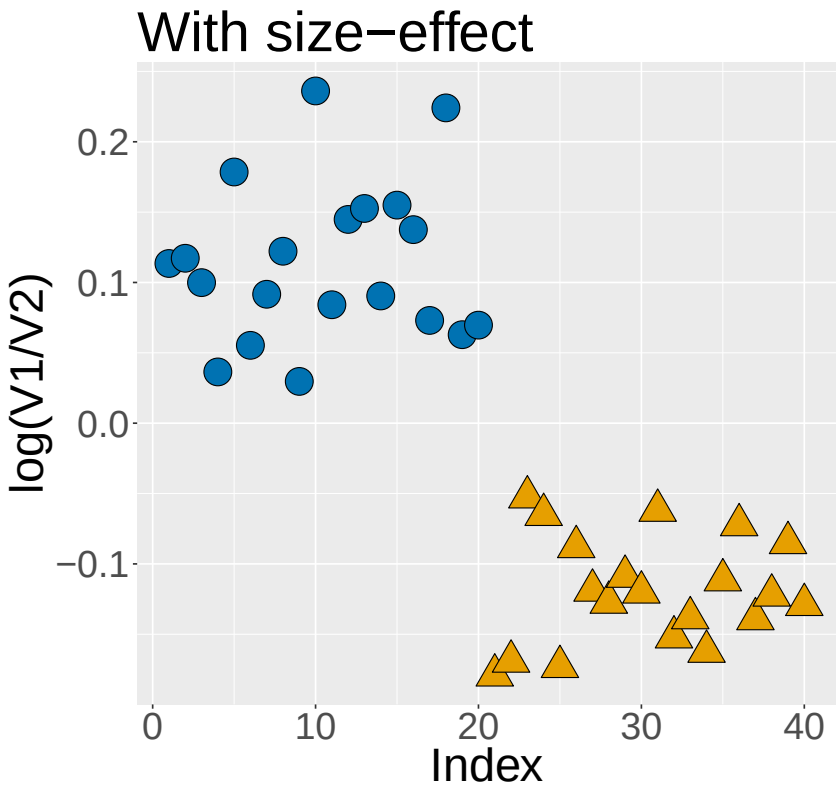
Size-effect


Without size−effect
value
50
45
40
Group
Control
Disease
V1
V2
V3
V4
variable


With size−effect
value
100
50
V1
V2
V3
V4
variable

Size-effect: Use of log-ratios





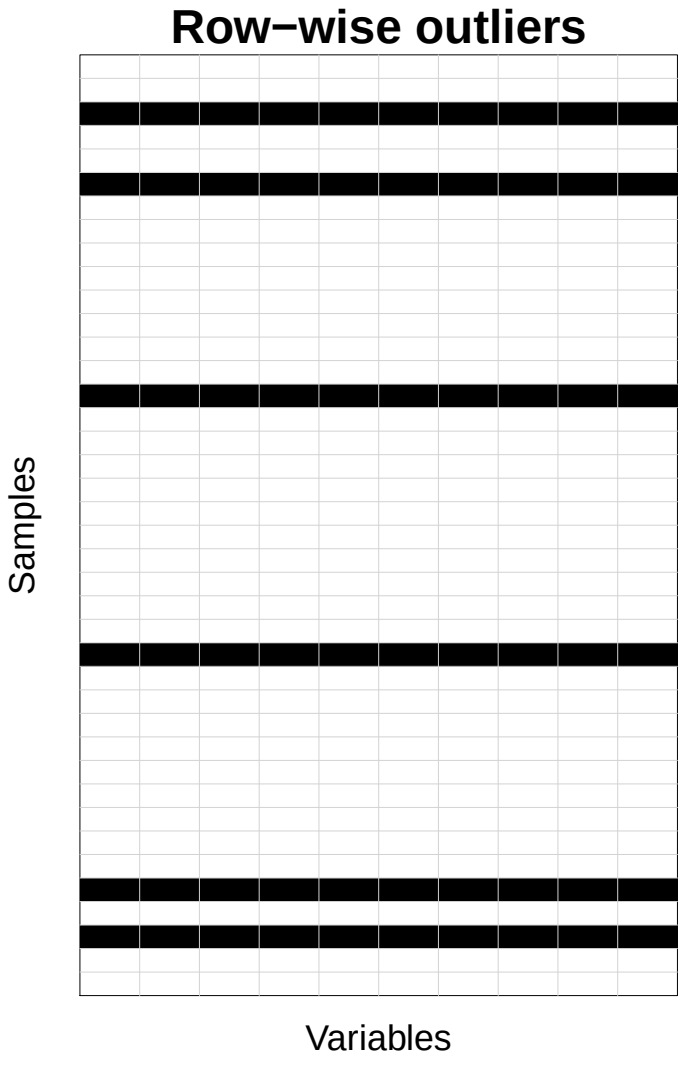
Title - second part

The use of log-ratio methodology in **cell-wise outlier diagnostics**

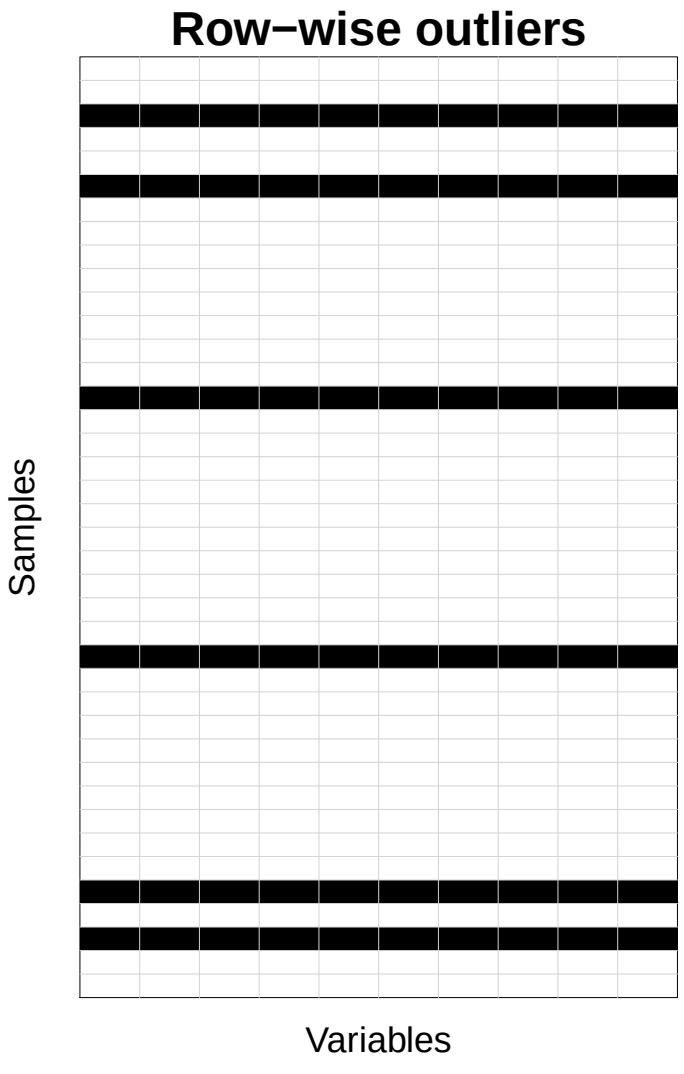
Title - second part

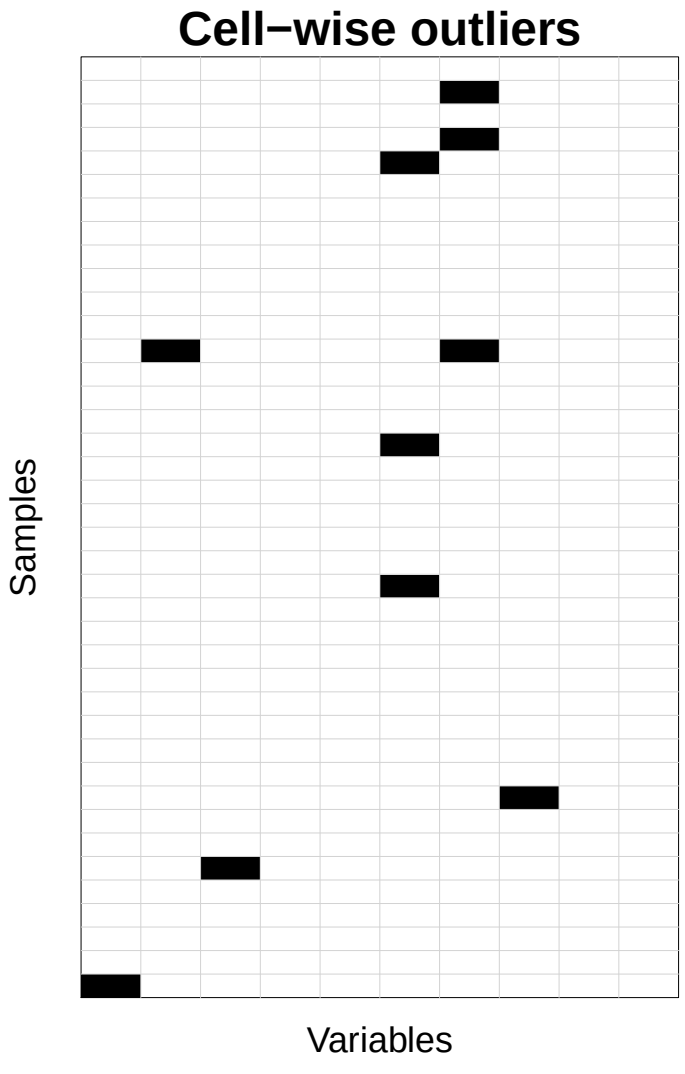
The use of log-ratio methodology in **cell-wise outlier diagnostics**

Peter J. Rousseeuw, Wannes Van Den Bossche. "Detecting deviating data cells." Technometrics 60.2 (2018): 135-145.

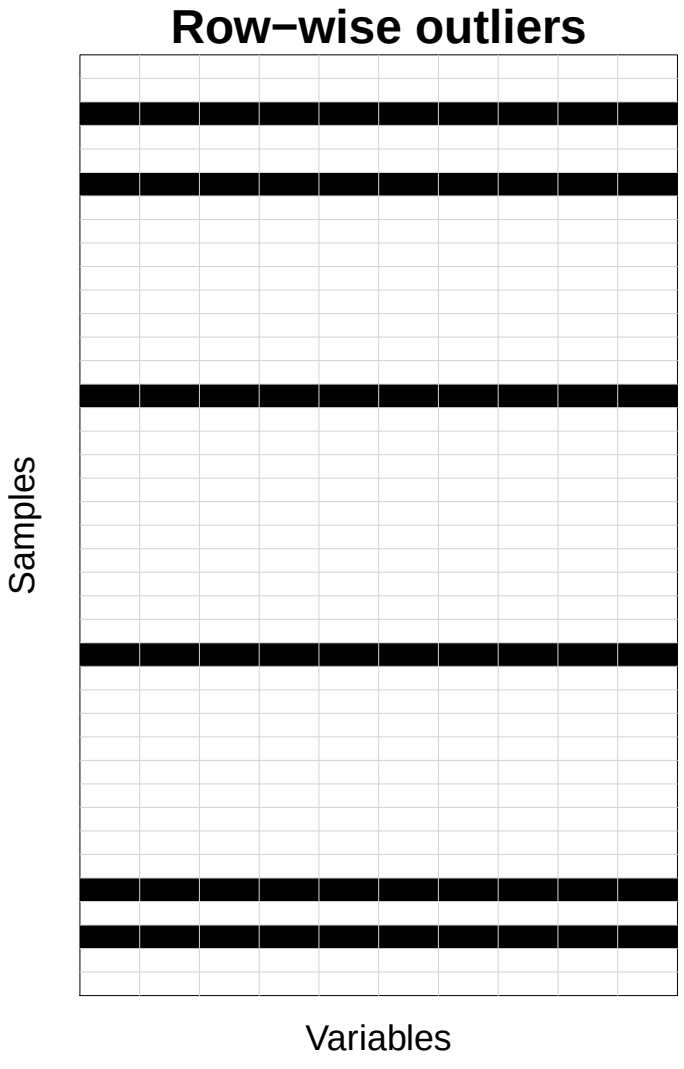

Row−wise outliers
Samples
Variables

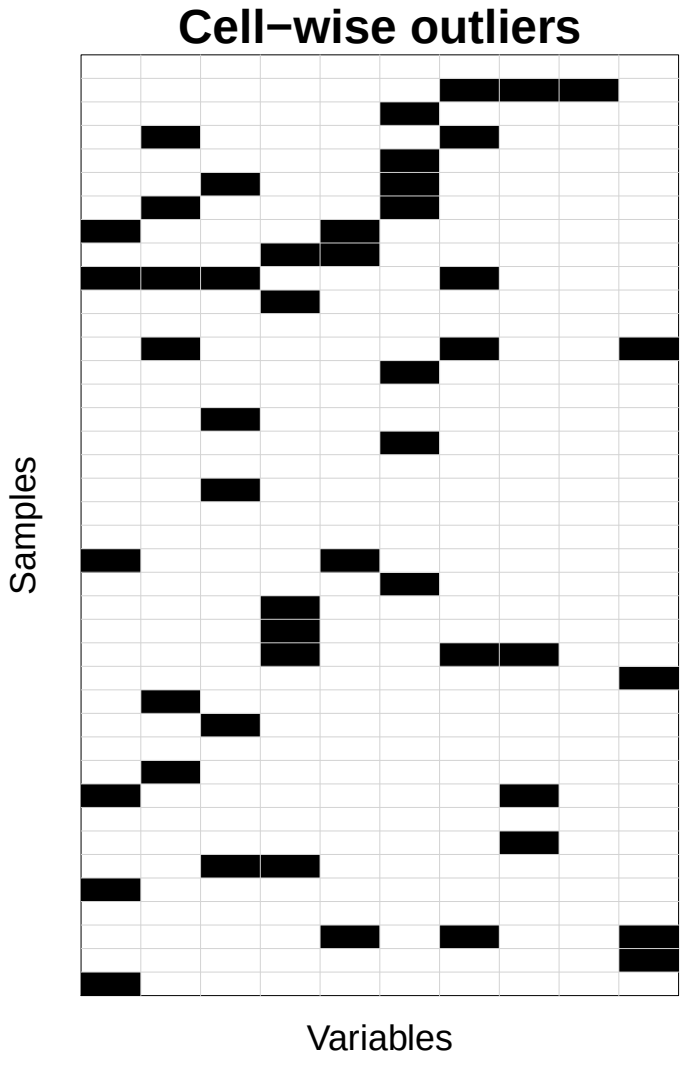
Outlier diagnostics





Outlier diagnostics





Cell-wise rPLR (1)

Method:

Cell-wise rPLR

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 $j, k = 1, \ldots, D$

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

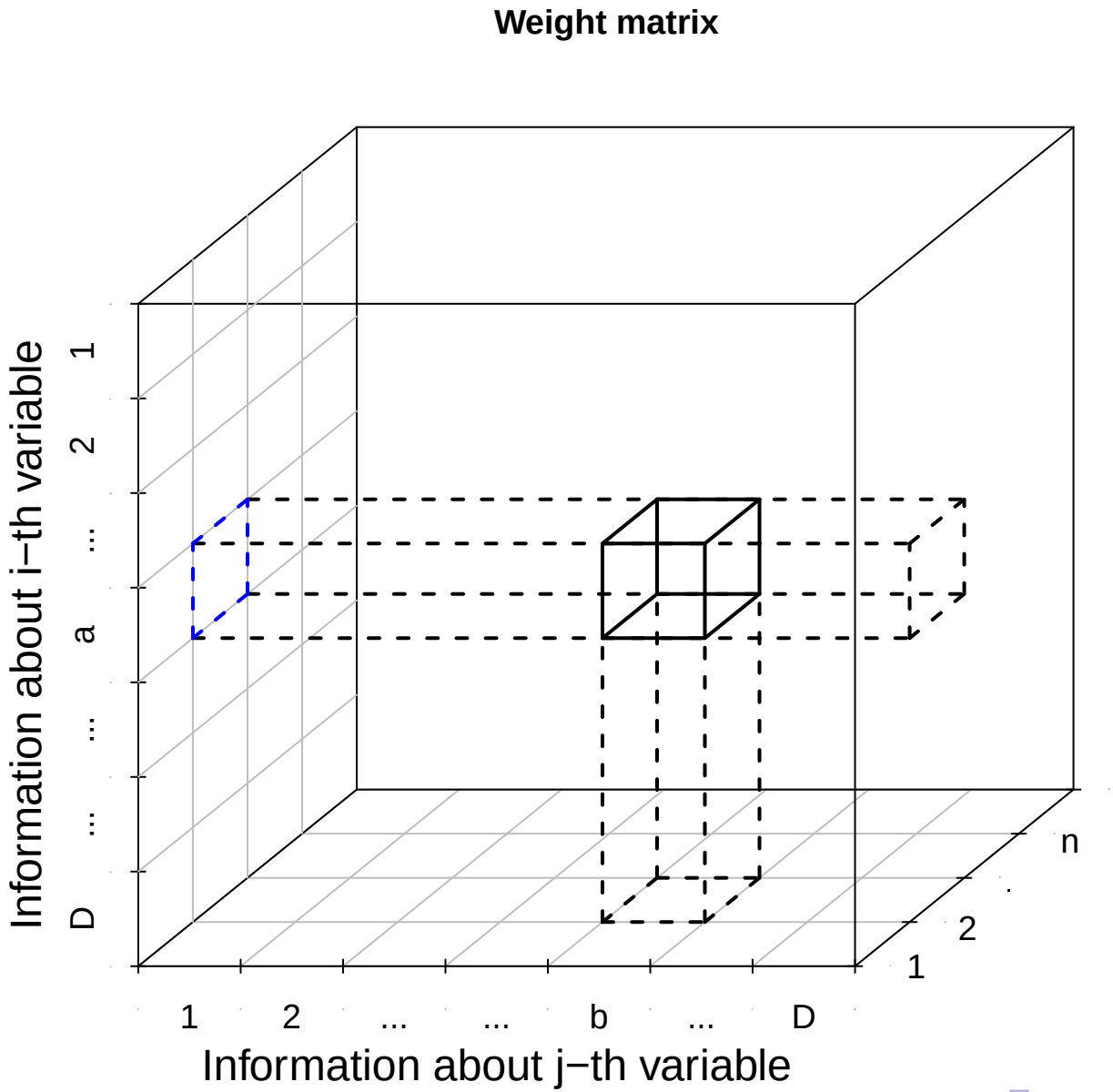
- where *var* denotes a variance
- Elements inside *var* () will be robustly centred + scaled and each will get a weight
- There are in total $\frac{D \times (D-1)}{2}$ possible log-ratios

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

Three-way weight matrix

- All weights can be stored in the three-way matrices W, with D rows, D columns and n slices
- Information about observations is in certain slices
- Information about cells in rows of slices

Projection to original dimensions



Diagnostics

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

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

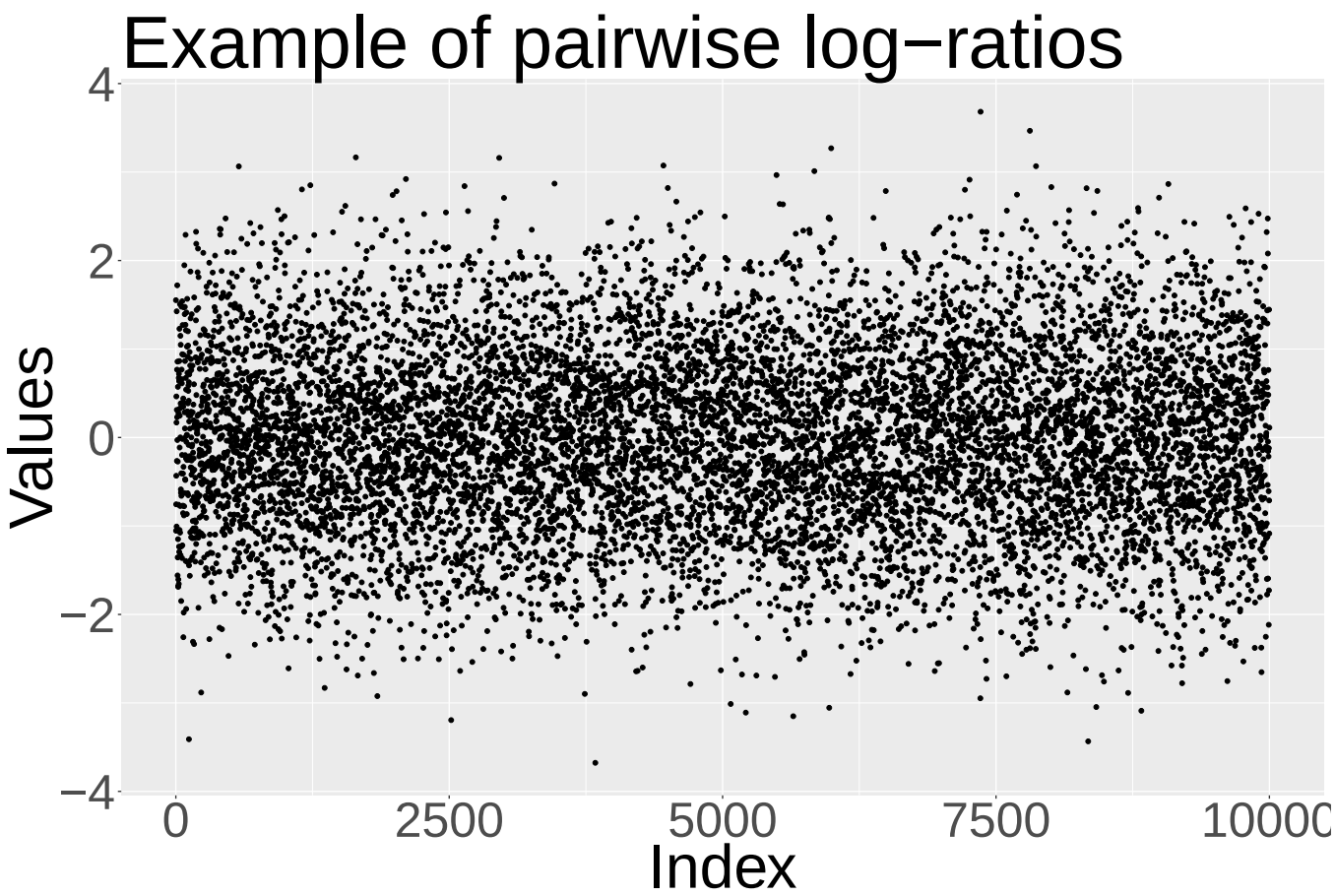
- Or Median

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

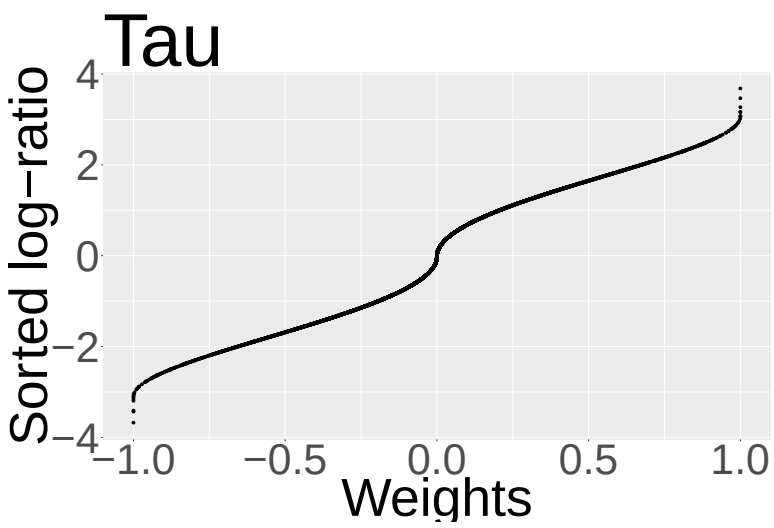
Applying weighting function

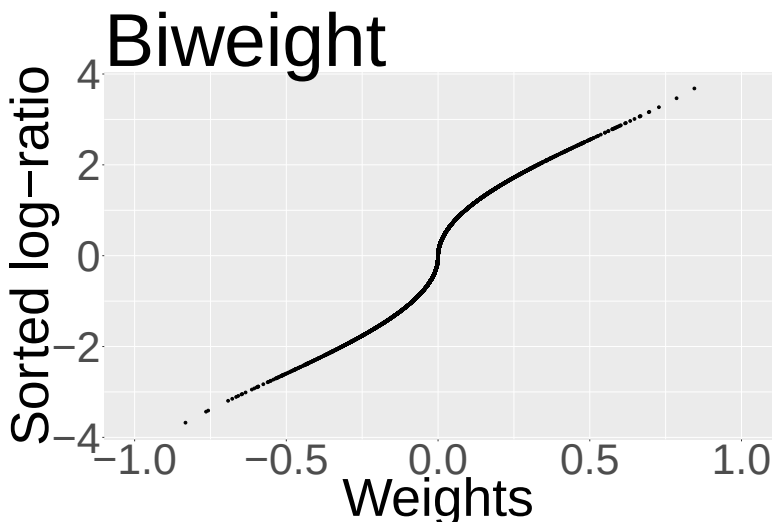
Weighting functions:

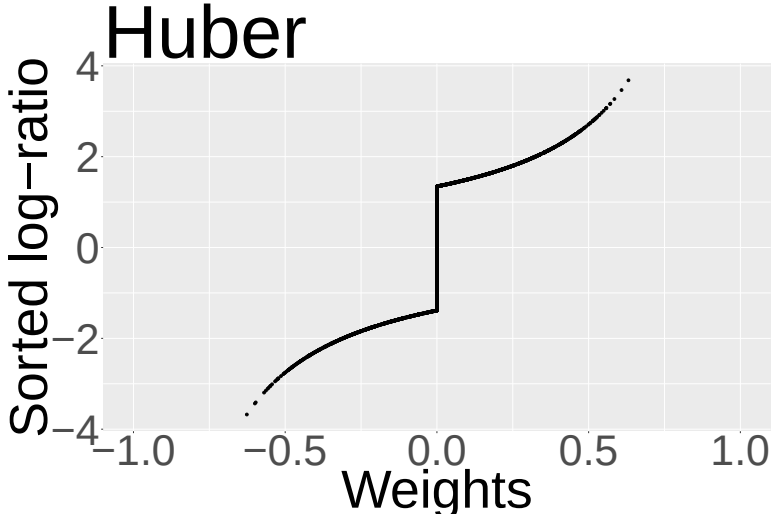
- τ function
- Tukey's biweight function
- Huber function
- Hampel function

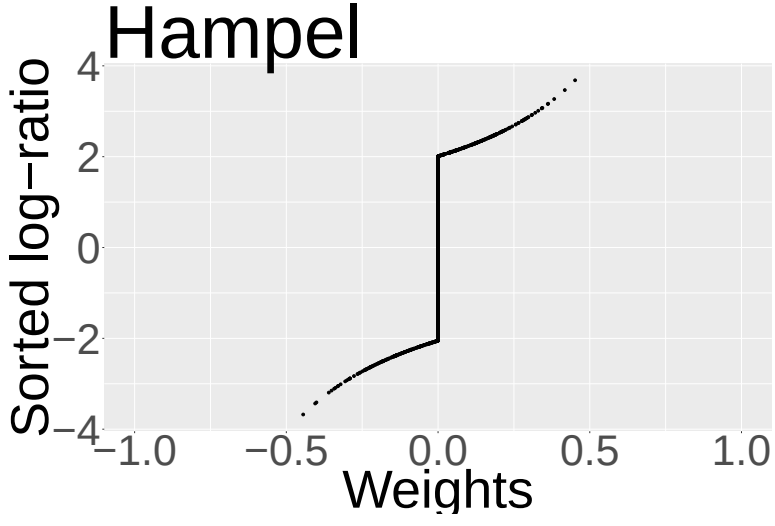

Example of pairwise log−ratios
Values
4
2
0
−2
−4
0
2500
5000
7500
10000
Index

Cell-wise outliers



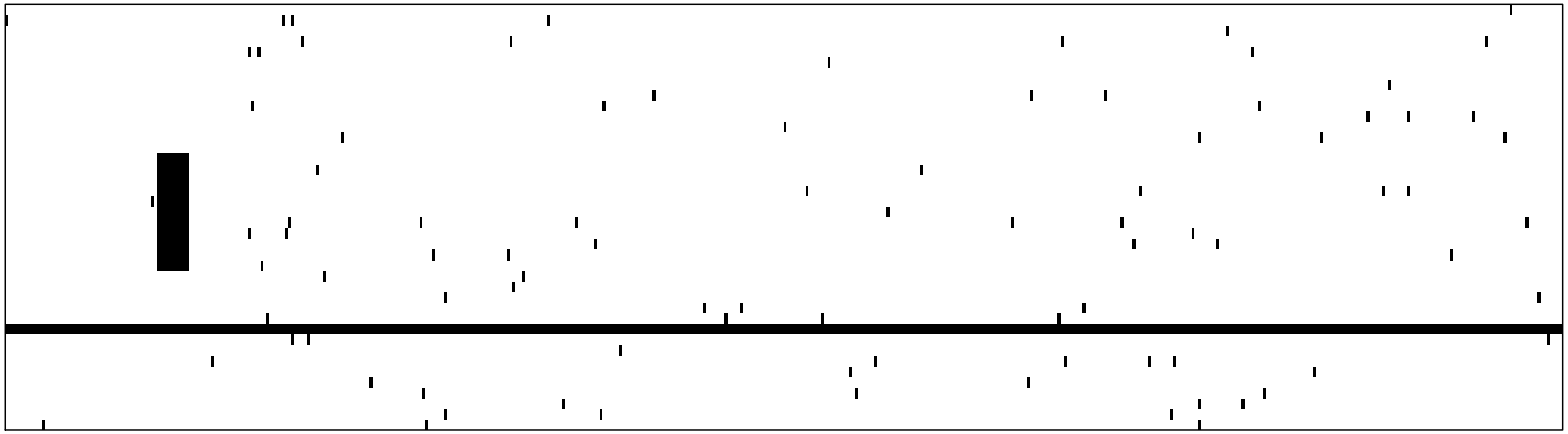




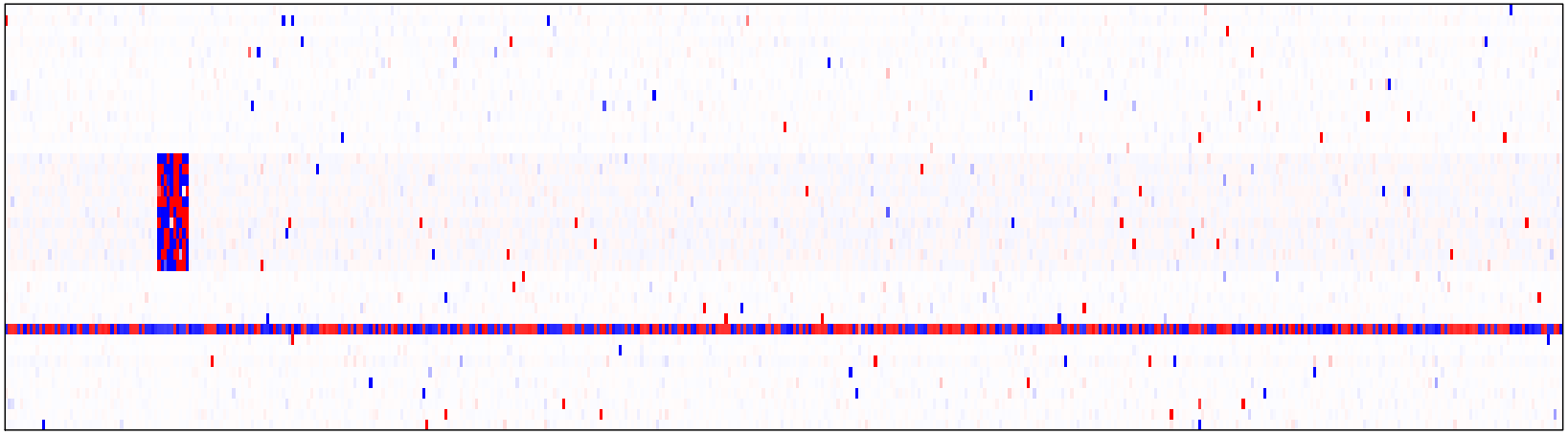


Diagnostics example: τ

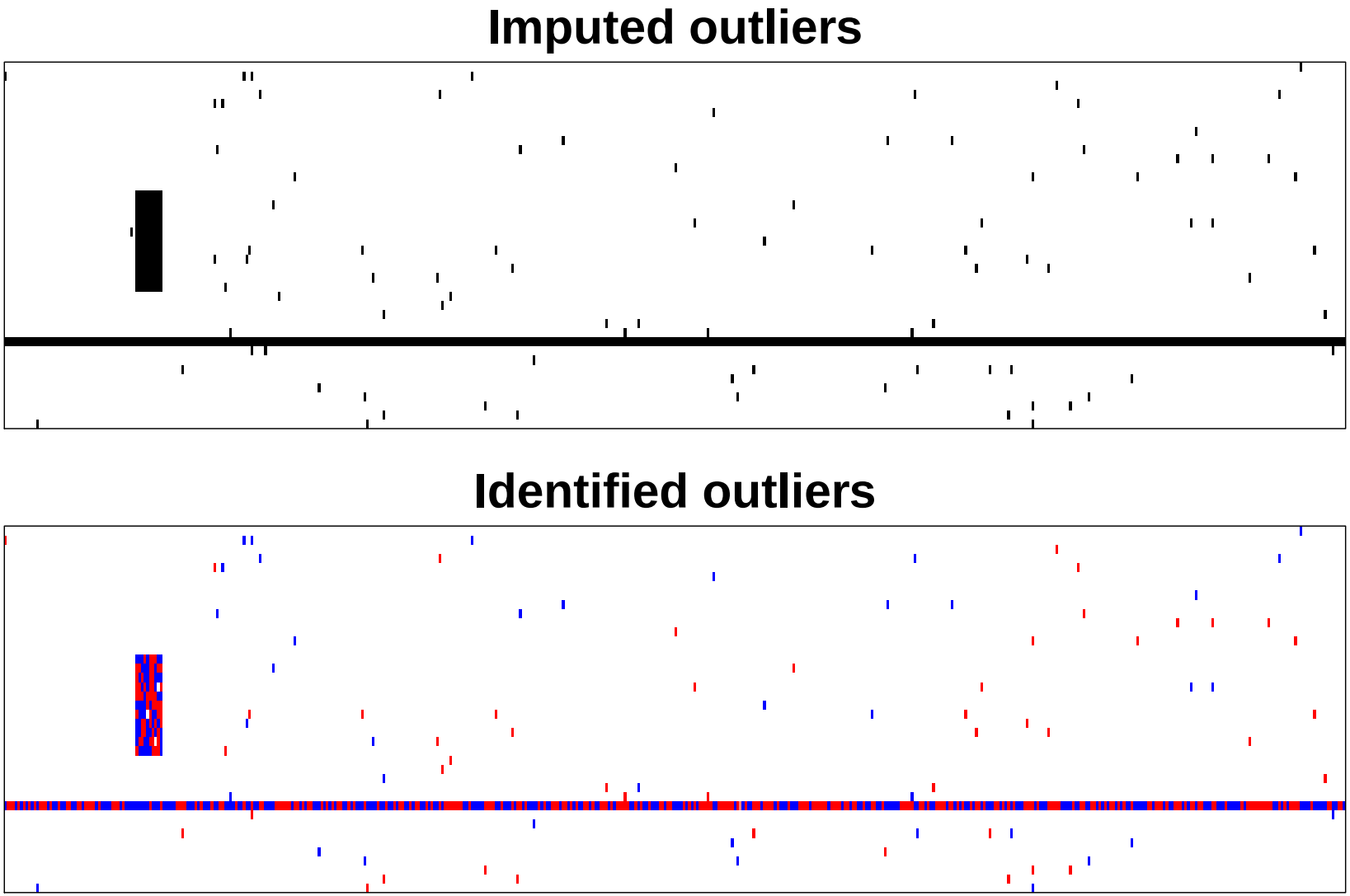
Imputed outliers



Identified outliers

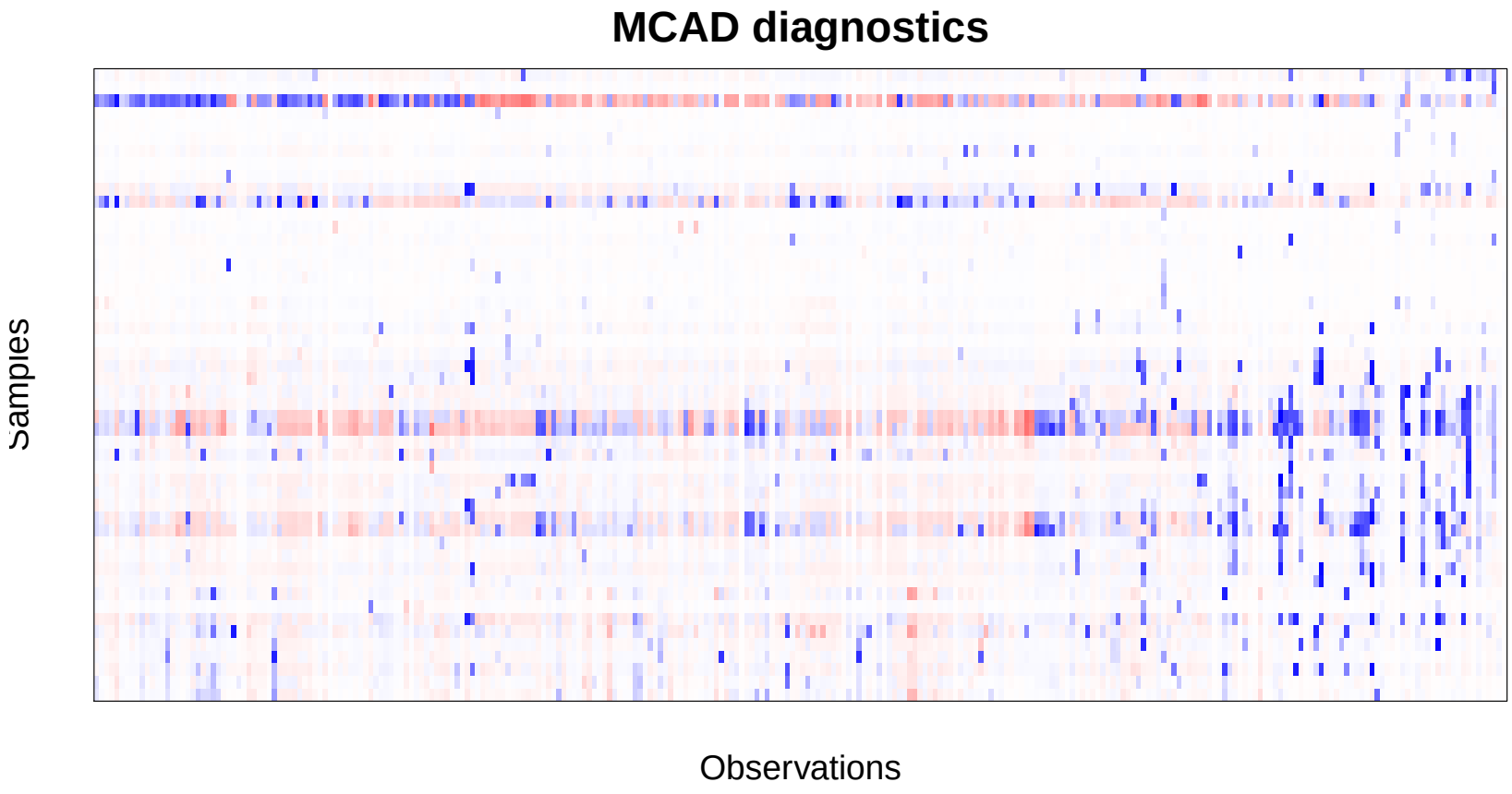


Diagnostics example: Hampel



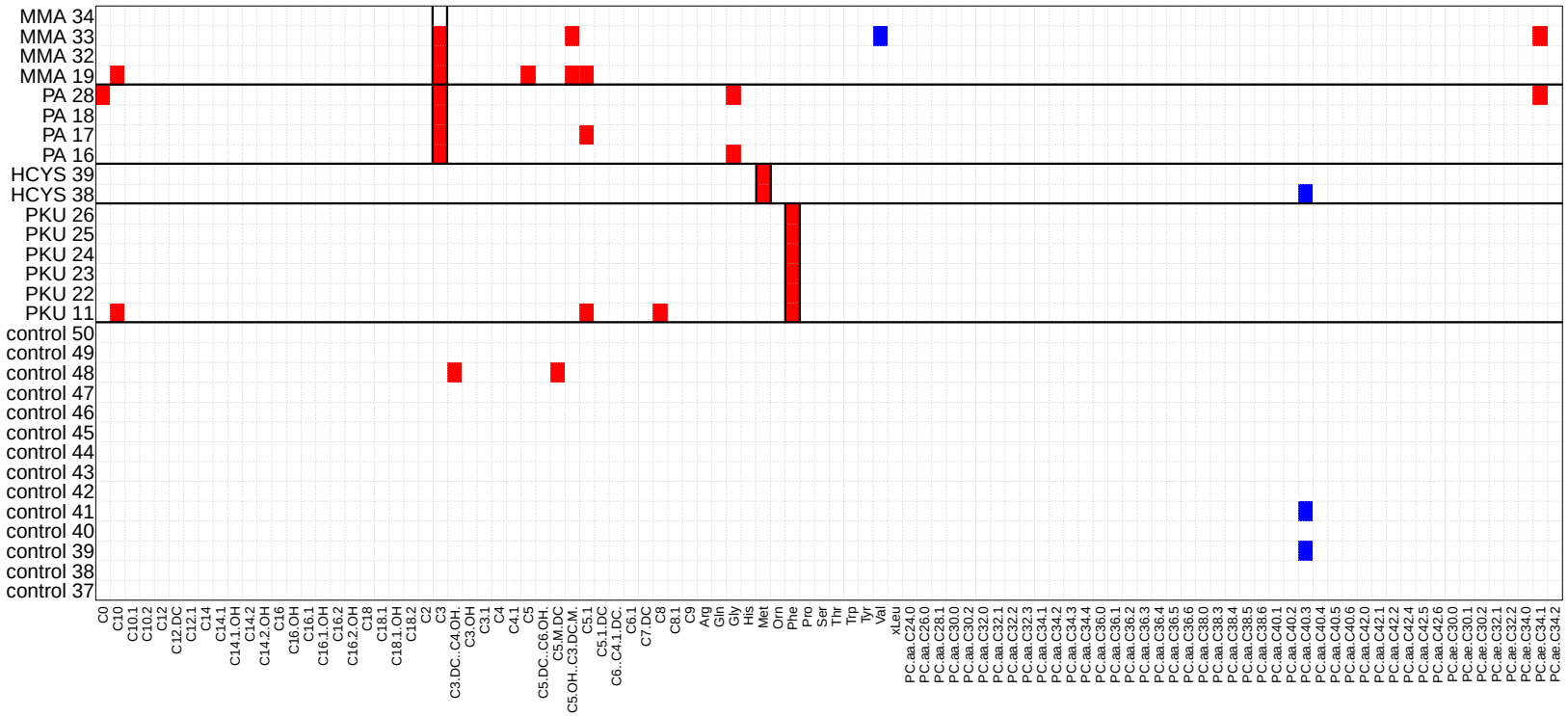
Metabolomic dataset MCADD

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



Biomarkers: Hampel

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



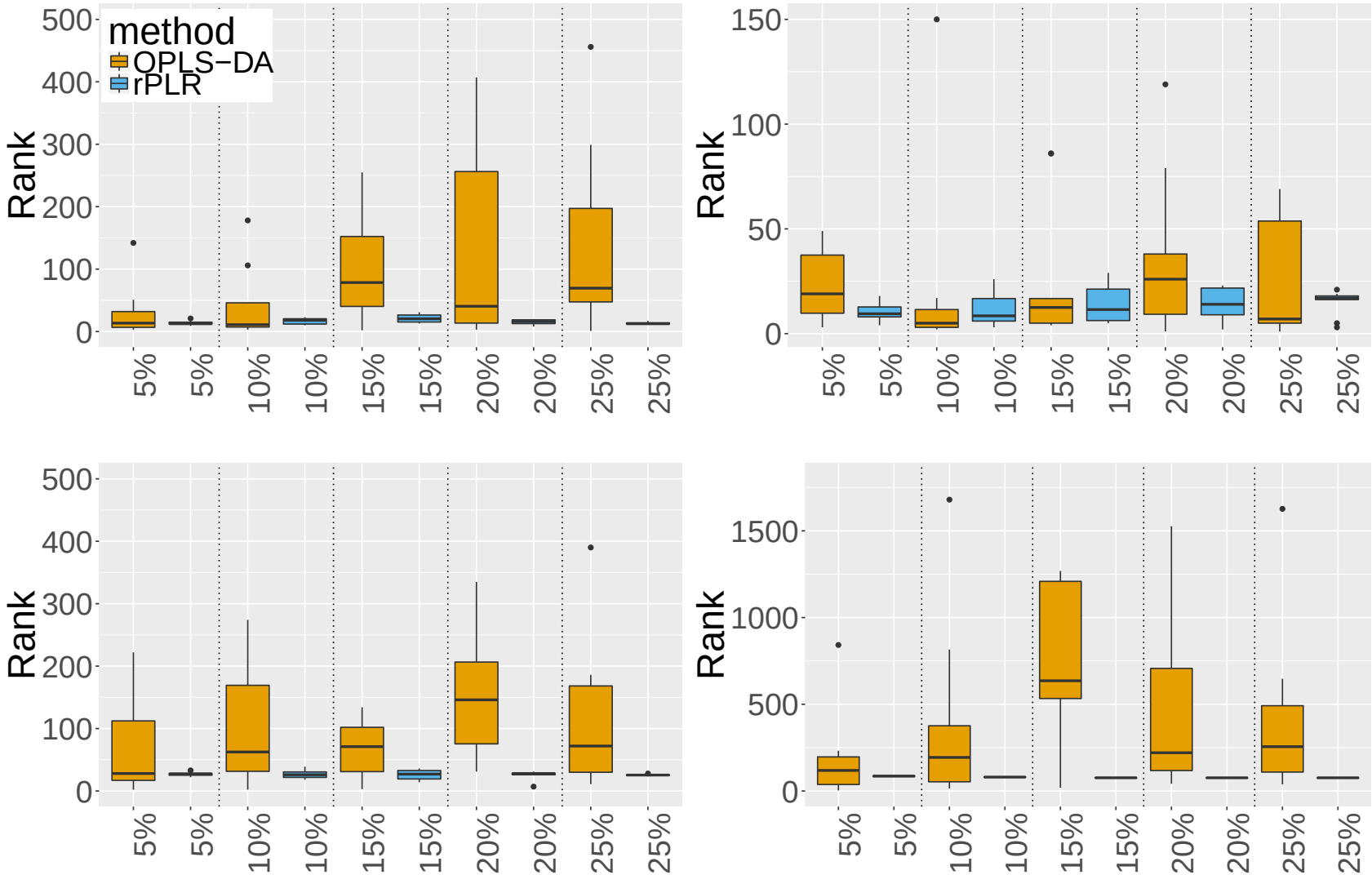
Biomarker ranking

- Let's consider two group problem with n_1, n_2 samples
- For each variable j:

$$V_j = |median_{i=1,\ldots,n_1} m_{ij} - median_{i=1,\ldots,n_2} m_{ij}| \qquad (4)$$

- Permutation test can be used in order to set a cut-off values

Metabolomic dataset



R-implementation

Cell-wise outliers

Threshold: 0.45

Weighting function: Hampel

Aggregation: Median

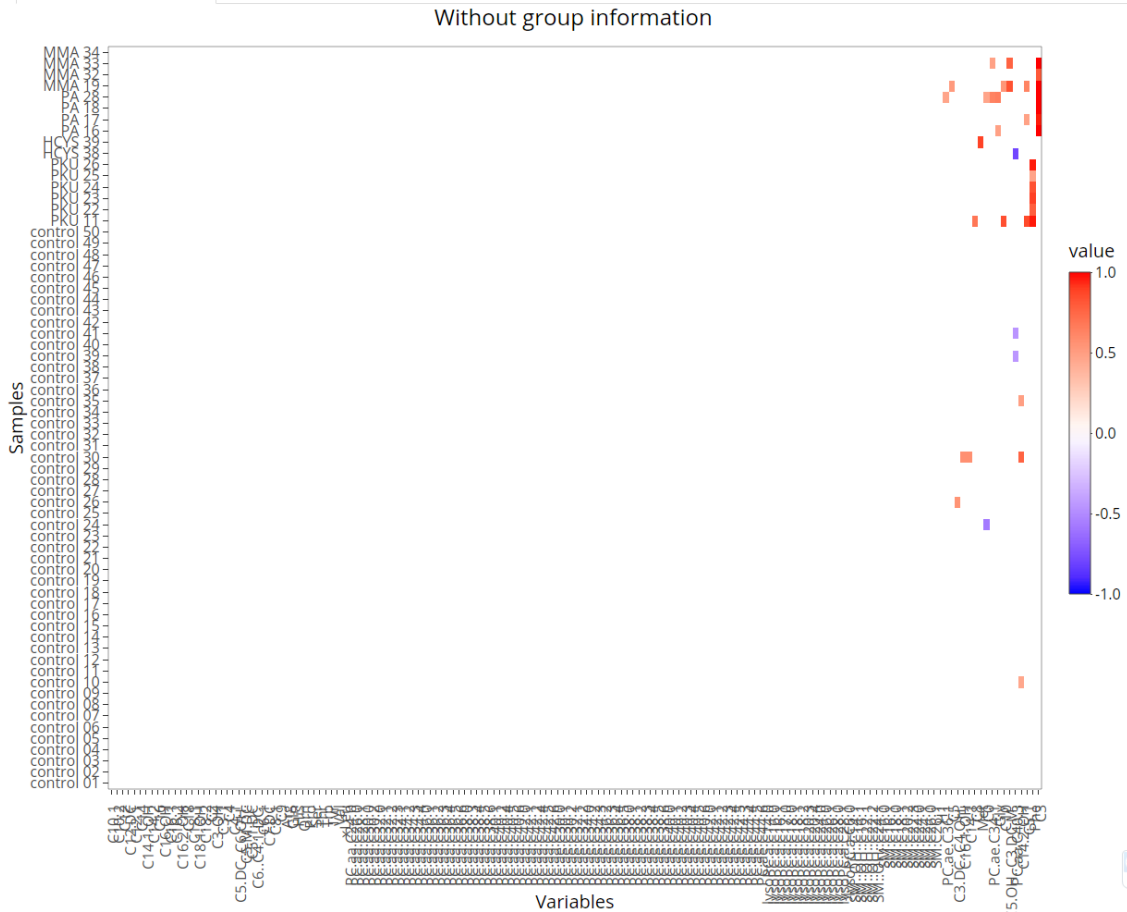
Sorting: By Number bigger than zero

Sort according to: 1 – 66

Filename

Download Plot

Without group information | With group information



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.


`https://github.com/walachja/Cell-wise_rPLR`