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

Jan Walach\textsuperscript{1}, P.Filzmoser\textsuperscript{1}, K.Hron\textsuperscript{2}

\textsuperscript{1}TU Wien, \textsuperscript{2}Palacky University Olomouc

June 11, 2018
# Outline of DSSV 2018 presentation

def dataset_directory()
    setwd("/home/jan/Dropbox/Git/Batch2/Prezentace")
    set.seed(1314233448)
    for (i in 1:7)
        {
            p <- sample(c('Talking', 'Talking', 'Describing a plot', 'Getting lost in a complicated formulas'), 3)
            print(p)
            for (j in 1:length(dir(paste0(getwd(),'/',p))))
            {
                temp = list.files(path = paste0(getwd(),'/',p), pattern="*")
                mySlides = lapply(temp, read.delim)
            }
        }

[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"
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

Size-effect

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

Without size–effect

With size–effect
Size-effect: Use of log-ratios

With size-effect

V1 vs. V2

log(V1/V2) vs. Index
The use of log-ratio methodology in cell-wise outlier diagnostics
The use of log-ratio methodology in **cell-wise outlier diagnostics**

# Outlier diagnostics

## Row-wise outliers

<table>
<thead>
<tr>
<th>Samples</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outlier diagnostics

Row-wise outliers

Variables
Samples

Cell-wise outliers

Variables
Samples
Outlier diagnostics

Row-wise outliers

Cell-wise outliers

Samples

Variables

Samples

Variables
Method:
Cell-wise rPLR

Steps:

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

Data

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

Based on idea of robust variation matrix

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

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

- where \( \text{var} \) denotes a variance

- Elements inside \( \text{var} () \) will be robustly centred $\pm$ scaled and each will get a weight

- There are in total $\frac{D \times (D-1)}{2}$ possible log-ratios
Computing log-ratios

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

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

Weight matrix

Information about $i$-th variable

Information about $j$-th variable

1 2 ... ... b ... D
D ... ... a ... 2 1
1 2 n
Diagnostics

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

\[ m_{ij} = \frac{1}{D} \sum_{k=1}^{D} w_{jki} \]  \hspace{1cm} (2)

- Or Median

\[ m_{ij} = \text{median}_{k}(w_{jki}) \]  \hspace{1cm} (3)
Applying weighting function

Weighting functions:

- $\tau$ function
- Tukey’s biweight function
- Huber function
- Hampel function
Example of pairwise log-ratios
Cell-wise outliers

Tau

Biweight

Huber

Hampel

Sorted log-ratio vs. Weights
Diagnostics example: τ

- Imputed outliers
- Identified outliers
Diagnostics example: Hampel

Imputed outliers

Identified outliers
• 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.
Let’s consider two group problem with $n_1$, $n_2$ samples

For each variable $j$:

$$V_j = |\text{median}_{i=1,\ldots,n_1} m_{ij} - \text{median}_{i=1,\ldots,n_2} m_{ij}|$$  \hspace{1cm} (4)

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

The diagrams illustrate the rank of different methods (OPLS-DA, rPLR) across various percentage thresholds (5%, 10%, 15%, 20%, 25%). The box plots show the distribution of ranks for each method at different thresholds, with median, quartiles, and outliers indicated.
R-implementation

Cell-wise outliers

Threshold:

Weighting function:

Aggregation:

Sorting:

Sort according to:

Filename

Download Plot

Without group information
References


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