Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios

Jan Walach1 | Peter Filzmoser1 | Štěpán Kouřil2,3 | David Friedecký2,3 | Tomáš Adam2,3

1 Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna Austria
2 Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc Czech Republic
3 Department of Clinical Biochemistry, University Hospital Olomouc, Olomouc Czech Republic

Correspondence
Peter Filzmoser, Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria.
Email: P.Filzmoser@tuwien.ac.at

Funding information
Austrian Science Fund; Grantová Agentura České Republiky, Grant/Award Number: 18-12204S I1910-N26(15-34613L)

Data outliers can carry very valuable information and might be most informative for the interpretation. Nevertheless, they are often neglected. An algorithm called cellwise outlier diagnostics using robust pairwise log ratios (cell-rPLR) for the identification of outliers in single cell of a data matrix is proposed. The algorithm is designed for metabolomic data, where due to the size effect, the measured values are not directly comparable. Pairwise log ratios between the variable values form the elemental information for the algorithm, and the aggregation of appropriate outlyingness values results in outlyingness information. A further feature of cell-rPLR is that it is useful for biomarker identification, particularly in the presence of cellwise outliers. Real data examples and simulation studies underline the good performance of this algorithm in comparison with alternative methods.

KEYWORDS
biomarker, cell-rPLR, cellwise outliers, log ratio, metabolomics, robust method

1 INTRODUCTION

Metabolomic data, as well as many other data sets from “omics” disciplines, are high dimensional, with many variables and commonly limited by few observations, originating from two or more different groups (controls and diseased). The groups can typically be distinguished at the basis of few variables only, the so-called biomarkers. Once they are identified, they are important for the interpretation of the group differences.1,2

Biomarker identification is getting more challenging if outliers are present in the data.3 Outliers in this context can be observations that are somewhat different in their data structure compared with the data majority, and this difference may be caused by measurement problems, different data preprocessing, inconsistencies among the observations, etc.4,5 An outlying observation does not necessarily differ in all the variable values, but it could differ just in few variables. This difference could be a data artifact, but it could also refer to a biomarker, for which a difference is to be expected. This means that for the purpose of biomarker identification, outliers could be disturbing if they are related to data artifacts or even helpful otherwise. In the latter case, one would expect that the outliers form a pattern; ie, all observations from that group should have outlying values for the respective biomarker.
Traditionally, outlier identification has been carried out “rowwise,” assuming that the observations are arranged in the rows of the data matrix. This means that if a method identifies an outlier, the complete observation is flagged as such. This situation is visualized in Figure 1 (left), which shows the cells of a data matrix, and the dark cells refer to outliers. Robust statistical estimators would then typically downweight outlying observations. In contrast to that, Figure 1 (right) refers to a scheme of “cellwise” outliers, where single cell of the data matrix (colored in black) is identified as an outlier. Thus, for each observation, different variables can be outlying. Especially for high-dimensional data, it might happen that most of the observations will contain at least one cellwise outlier. It would not make much sense to downweight those observations, which contain an outlying cell, since most of the observations would then get downweighted. Cellwise outlier detection is a quite recent topic in robust statistics, as well as the development of robust estimators with cellwise outliers. In fact, since our proposed algorithm will be based on variable pairs, there is some similarity to the algorithm of Rousseeuw and Bossche.

Data from metabolomics often have big differences in abundance of individually measured variables. Whereas some variables have low abundances, others are on a much larger scale. This situation is undesirable, since statistical methods could be biased toward higher level abundances. Furthermore, many statistical methods assume that the errors fluctuate around 0 with constant variation. Unfortunately, often with increasing abundances, the variation of noise also increases. Such a situation is called heteroscedasticity. A further important characteristic of metabolomic data is the so-called size effect. This refers to a situation in which the concentration or abundance is generally different for each sample in the data set, e.g., in the analysis of urine samples where concentration strongly depends on the water intake of the patient and other factors. Thus, the obtained data values are not directly comparable, and the data need to be preprocessed first before applying a statistical method. Prepossessing can be done by making use of a specific data transformation and normalization, e.g., the total sum normalization (TSN) or the probabilistic quotient normalization (PQN). An alternative is to use the log-ratio methodology from compositional data analysis, which is based on pairwise log ratios. Since for two observations, \( x \) and \( y \), and any positive constant \( s \) (representing the size effect), \( \ln(\frac{s \cdot x}{s \cdot y}) = \ln(\frac{x}{y}) \), preprocessing is not necessary here.

In this paper, we will introduce an algorithm called cell-rPLR. The goal of cell-rPLR is twofold: It can be used for (a) cellwise outlier identification and for (b) biomarker identification. Section 2 describes the theoretical basis of the method. Section 3 introduces a diagnostics plot for cellwise outlier identification. In a simulation scenario, cell-rPLR is compared with an alternative approach for cellwise outlier detection. Section 4 shows how cell-rPLR is used for biomarker identification, and the performance is evaluated in scenarios where cellwise outliers are artificially included in the data. Section 5 summarizes the paper, provides information about software, and concludes.

2 | METHOD

Let us assume a data set arranged in a data matrix \( X \), with \( n \) samples and \( d \) variables. The matrix \( X \) consists of elements \( x_{ij} \), where \( i = 1, \ldots, n \) and \( j = 1, \ldots, d \). There are \( G \geq 2 \) groups of samples in our data, and one can rearrange the samples so that...
samples belonging to one group are gathered together in one block in \( \mathbf{X} \). Each block is denoted as \( \mathbf{X}^g \) with elements \( x_{ij}^g \) for \( i=1, \ldots, n_g, j=1, \ldots, d \) and \( g=1, \ldots, G \) and \( n_1 + \cdots + n_g = n \).

The proposed method cell-rPLR consists of three main steps. In the first step, we use the information of the log ratios between pairs of variables. In the second step, the log ratios are robustly centered and scaled, and an outlyingness function is applied. Finally, the third step projects the data to the original dimensions \( n \times d \). In the following, a detailed description of the individual steps is provided.

### 2.1 Centered and scaled pairwise log ratios

Consider for a pair of variables, with index \( j,k \in \{1, \ldots, d\} \), the log ratios of their observations:

\[
\ln \left( \frac{x_{1j}^{(1)}}{x_{1k}^{(1)}} \right), \ldots, \ln \left( \frac{x_{n1j}^{(1)}}{x_{n1k}^{(1)}} \right), \ln \left( \frac{x_{n1j+n2j}^{(2)}}{x_{n1j+n2k}^{(2)}} \right), \ldots, \ln \left( \frac{x_{Gnj}^{(G)}}{x_{Gnk}^{(G)}} \right).
\]

(1)

Clearly, the log ratios are 0 if \( j=k \), and exchanging denominator and nominator leads to the same log ratio but with different sign. Subsequently, we will assign an outlyingness value to each entry of the pairwise log ratios. In order to design an appropriate outlyingness function, the log ratios need to be centered and scaled first. Since potential group differences should not get lost, centering and scaling are performed based on the robust center and scale calculated for the majority group. In case that the group sizes of the biggest groups are equal, one can randomly select one of these biggest groups. For simplicity, suppose now that the first group is the biggest group, thus \( n_1 > n_g \), for \( 1 \leq g \leq G \). Further, we simplify the notation by defining \( y_{ijk} := \ln \left( \frac{x_{ij}^{(1)}}{x_{ik}^{(1)}} \right) \), for \( i=1, \ldots, n \), and \( j,k \in \{1, \ldots, d\} \). For the following steps, let us drop the indexes \( j \) and \( k \) for simplicity, and thus, \( y_i := y_{ijk} \). The log ratios of the first group are the values \( y_1, \ldots, y_{n_1} \).

Center and scale of the log ratios of the first group are estimated robustly\(^6\) as

\[
\bar{y}_1 = \frac{\sum_{i=1}^{n_1} v_i y_i}{\sum_{i=1}^{n_1} v_i},
\]

(2)

where

\[ v_i = \omega \left( \frac{y_i - \text{median}(y_1, \ldots, y_{n_1})}{s_1} \right), \]

(3)

and \( s_1 = \text{MAD}(y_1, \ldots, y_{n_1}) \) is the median absolute deviation, defined as

\[
\text{MAD}(y_1, \ldots, y_{n_1}) = 1.483 \cdot \text{median}_i(|y_i - \text{median}_j(y_j)|).
\]

(4)

The function \( \omega(\cdot) \) in Equation (3) is Tukey’s biweight function\(^{16}\), defined as

\[
\omega_c(u) = \left( 1 - \left( \frac{u}{c} \right)^2 \right)^2 \cdot I(u, c),
\]

(5)

with

\[
I(u, c) = \begin{cases} 1, & \text{for } |u| < c, \\ 0, & \text{otherwise}. \end{cases}
\]

(6)

The tuning constant is usually chosen as \( c=4.685 \). For more details, we refer to Yohai and Zamar\(^{17}\) and Maronna and Zamar,\(^{18}\) who introduced these concepts in the framework of robust scale estimation.
The robustly centered and scaled values are obtained as
\[
\tilde{y}_i = y_i - \frac{y_1}{s_1} \quad \text{for } i = 1, \ldots, n.
\] (7)

Centering and scaling are done for fixed indexes \(j,k \in \{1,\ldots,d\}\), and now, going back to the notation including these indexes, we end up with robustly centered and scaled values \(\tilde{y}_{ijk}\). Note that for \(j=k\), the function arguments in (3) are not defined, because \(s_1\) would be 0. We will thus set the values \(\tilde{y}_{ijk} = 0\) whenever \(j=k\). Further, one can see that \(\tilde{y}_{ijk} = -\tilde{y}_{ikj}\), and therefore, it is sufficient to actually compute only the values \(\tilde{y}_{ijk}\) for \(j<k\), which saves computational effort.

### 2.2 Outlyingness functions

The robustly centered and scaled pairwise log ratios contain information about outlyingness, and this information will be revealed by applying an appropriate outlyingness function to these values. An outlyingness function as proposed in Equation (5) is blue commonly used in robust statistics, eg, in the regression context to downweight (absolute) large residuals, in order to obtain a robust estimator. However, it would not be appropriate for the outlier detection task, since the resulting outlyingness values are in the interval \([0,1]\), and one would lose the sign information of the log ratios. This information will be important, because positive values would refer to a dominance of the nominator and negative values to a dominance of the denominator. Therefore, we propose the adjusted Tukey biweight function as
\[
\omega_c^*(u) = \omega_c(u) \cdot \text{sgn}(-u) + \text{sgn}(u),
\] (8)

with the sign function
\[
\text{sgn}(v) = \begin{cases} 
1, & \text{for } v \geq 0, \\
-1, & \text{otherwise}, 
\end{cases}
\] (9)

yielding values in \([-1,1]\). Figure 2 shows the original definition of the Tukey biweight weights (left plot) and compared with the adjusted version leading to outlyingness values (right plot).

The adjusted outlyingness function is applied to the robustly centered and scaled pairwise log ratios from (7). Outlyingness values around 0 represent nonoutliers, and values closer to \(-1\) or \(+1\) represent potential outliers.

The shape of the outlyingness function will determine the characteristics of the outlier detection method, and thus, other outlyingness functions shall be considered as well. In the literature on robust statistics, many proposals are available, such as Huber's\(^4\) and Hampel's\(^19\) functions. Below, we will propose the adjusted versions, resulting in outlyingness values in the interval \([-1,1]\).

Huber function: The original definition is
\[
\omega_k(u) = \min \left(1, \frac{k}{|u|}\right),
\] (10)
where \( k \) is a tuning parameter, typically taken as 1.345.\textsuperscript{4} The assignment of 1 to a broader range of values improves the efficiency of robust estimators, while still keeping their robust properties. The adjusted version

\[
\omega^*_k(u) = (\omega_k(u) - 1) \cdot \text{sgn}(-u),
\]

assigns outlyingness values of 0 to those (nonoutlying) log ratios, which are still in the usual range. The resulting shape of the adjusted Huber function can be seen in Figure 2 (right); the left plot shows the original definition.

Hampel function: The original definition is

\[
\omega_h(u) = \begin{cases} 
1 & |u| \leq c_1 \\
\frac{c_1}{|u|} & c_1 < |u| \leq c_2 \\
\frac{c_2 - |u|}{c_1} & c_2 < |u| \leq c_3 \\
\frac{c_3 - c_2}{|u|} & c_3 < |u| \\
\end{cases},
\]

where the tuning parameters are typically chosen as \( c_1 = z_{0.95} \), \( c_2 = z_{0.975} \), and \( c_3 = z_{0.99} \), with \( z_q \) as the \( q \) quantile of the standard normal distribution. The Huber function approaches 0 asymptotically, whereas in the Hampel function, one obtains outlyingness values of 0 according to the tuning parameter \( c_3 \). The adjusted version of the Hampel function is

\[
\omega^*_h(u) = (\omega_h(u) - 1) \cdot \text{sgn}(-u),
\]

which again provides values in \([-1,1]\). Figure 2 shows the original (left) and adjusted (right) Hampel function.

### 2.3 Aggregation of outlyingness values

Let us denote the adjusted outlyingness function by \( \omega^*(\cdot) \), which refers to one of the proposed functions in Section 2.2. We apply this function to the centered and scaled pairwise log ratios (see Section 2.1), resulting in outlyingness values

\[
w^*_{ijk} = \omega^*(\tilde{y}_{ijk}),
\]

for \( i = 1, \ldots, n \) and \( j, k \in \{1, \ldots, d\} \). These outlyingness values are stored in an array \( W^* \) with \( n \) rows, \( d \) columns, and \( d \) slices.

Since we aim at a method for cellwise outlier detection, the outlyingness values in the array \( W^* \) need to be aggregated appropriately in order to obtain outlyingness values for each cell in the \( n \times d \) data matrix \( X \). For robustness reasons, we propose the aggregation into outlyingness values

\[
w_{ij} = \text{median}(w^*_{ij1}, w^*_{ij2}, \ldots, w^*_{ijd}),
\]

for \( i = 1, \ldots, n \) and \( j = 1, \ldots, d \), and they are collected in the \( n \times d \) outlyingness value matrix \( W \). Note that it would also be possible to aggregate the outlyingness values \( w^*_{ijk} \) according to the second index. This would result in the same values of aggregated outlyingness values but with reverse sign, because the considered outlyingness functions have the property \( \omega^*(u) = -\omega^*(-u) \) and because \( \tilde{y}_{ijk} = -\tilde{y}_{ikj} \).

### 2.4 Cell-rPLR algorithm for outlier diagnostics

The algorithm for cell-rPLR can be summarized as follows:

**Step 1:** Compute all pairwise log ratios \( \ln \left( \frac{X_{ij}^{(g)}}{X_{ik}^{(g)}} \right) \) for \( i = 1, \ldots, n \) and \( j, k \in \{1, \ldots, d\} \) with \( j > k \) (see 1).

**Step 2:** Center and scale them robustly according to the majority group. This gives values \( \tilde{y}_{ijk} \), for all \( i, j, k \).

**Step 3:** Apply an outlyingness function to \( \tilde{y}_{ijk} \), which yields outlyingness values \( w^*_{ijk} \) (see 14).

**Step 4:** Aggregate the outlyingness values according to (15) to obtain the final outlyingness values \( w_{ij} \), arranged in the outlyingness values matrix \( W \).
Step 3:Aggregate the values from the columns, and again, the results will be very similar to our proposed version. This would be computationally far too expensive, at least for high-dimensional data, and again, the results will be very similar to our proposed version.

Step 4: Compute the differences according to Equation (16), resulting in \( mj = \left| \text{median}(w^{(1)}_j) - \text{median}(w^{(2)}_j) \right| \). The larger the difference is, the more important the variable is for the discrimination of the groups. Note that the size of \( mj \) for different \( j \) can indeed be compared, since the outlyingness values are on the same scale. One can either sort the values \( mj \) in descending order or obtain a ranked variable list, with potential biomarkers at the beginning of the list. On the other hand, it might be desirable to obtain a cutoff value indicating potential biomarkers. Therefore, we will make use of a permutation test. Permutation tests\(^{20,21} \) are widely used for significance testing. They are based on resampling and try to estimate the distribution of the test statistic. The goal is to estimate a \( P \) value for the testing problem. In our case, the null hypothesis states that there is no difference between the two groups in the data for a certain variable, ie, \( mj=0 \), and thus, the variable is not a biomarker.

The permutations tests for cell-rPLR can be described as follows:

1. **Step 1**: Use as an input the matrix \( W_v \) with the elements \( w_{ijk} \), defined in (14).
2. **Step 2**: Randomly permute the values \( w_{ijk} \) according to the index \( j \), resulting in values \( w_{ijk}(b) \) for replication \( b \in \{1, \ldots, B\} \).
3. **Step 3**: Aggregate the values from the \( b \)th permutation as in Equation (15), yielding values \( w_j(b) \).
4. **Step 4**: Compute the differences according to Equation (16), resulting in \( mj(b) \), for \( j = 1, \ldots, d \).
5. **Step 5**: Compute the proportion
   \[
   \frac{1}{B} \sum_{b=1}^{B} (mj \leq mj(b)),
   \]
for \( j = 1, \ldots, d \), which is interpreted as \( P \) value for the \( j \)th variable. Here, \( mj \) refers to the values from (16) for the unpermuted data.

In our numerical experiments, we used \( B=1000 \). The computations are still feasible, because the input matrix \( W_v \) is fixed and with the \( B \) permutations, the \( P \) values for all variables are returned.

The permutations carried out in step 2 are done with respect to the index \( j \) only, which is numerically easy and quick to do. An alternative would be permuting all the elements in the blocks of the array \( W_v \). Technically, this would be more difficult and time-consuming, and the results would essentially be the same. The reason is because a permutation only in index \( j \) already destroys the group structure, since the original outlyingness values are based on values \( \hat{y}_{ijk} \), which are centered and scaled according to one of the groups. The median aggregation in (15) carries on this destroyed group structure to the test statistic (16).

Note that an alternative for the permutation test would be to start the permutations directly from the values \( \hat{y}_{ijk} \), ie, by permuting the log ratios already. This would be computationally far too expensive, at least for high-dimensional data, and again, the results will be very similar to our proposed version.

2.5 Cell-rPLR algorithm for biomarker identification

As noted above, the cell-rPLR can also be used for feature selection. In this case, however, this is limited only to the case of \( G=2 \) groups. The outlyingness values \( w_j \) from step 4 of the algorithm are arranged groupwise, and we have outlyingness values \( w^{(1)}_j = \{w_{1j}, \ldots, w_{nj}\} \) for the first group and outlyingness values \( w^{(2)}_j = \{w_{n+1,j}, \ldots, w_{n+1,}\} \) for the second group, for \( j = 1, \ldots, d \). For feature selection, we compare the medians in both sets of outlyingness values by

\[
 mj = \left| \text{median}(w^{(1)}_j) - \text{median}(w^{(2)}_j) \right|.
\]

The intensity of the color strongly depends on the chosen outlyingness function. Furthermore, one should keep in mind that the outliers are analyzed in comparison with one group (eg, the majority group) and need to be interpreted this way.

Note that an alternative for the permutation test would be to start the permutations directly from the values \( \hat{y}_{ijk} \), ie, by permuting the log ratios already. This would be computationally far too expensive, at least for high-dimensional data, and again, the results will be very similar to our proposed version.


3 | PERFORMANCE OF CELL-RPLR FOR OUTLIER IDENTIFICATION

The cell-rPLR algorithm results in outlyingness values $w_{ij}$, arranged in the outlyingness values matrix $W$; see Equation (15), which indicates cellwise outlyingness, and they can be visualized in a heatmap. A heatmap is a graphical visualization of the cells of a matrix, with the corresponding number of rows and columns, where each value of the data matrix is represented by a color information. We will use red color for positive outlyingness values and blue color for negative values, and the magnitude of the outlyingness value will determine the color saturation (values around 0 in light colors).

3.1 | Data sets

Three metabolomic data sets were used to demonstrate the usefulness of the method. The data sets differ in size and in the number of groups. For the last two data sets, expert knowledge about biomarkers is available.

**IMD:** This data set consists of plasma samples from infants (50 control samples and 16 samples) with different metabolic diseases analyzed in the Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Czech Republic. There are in total four different inherited metabolic disorders (IMD)—phenylketonuria (PKU), homocystinuria (HCYS), methylmalonic aciduria (MMA), and propionic aciduria (PA), each with different number of samples varying from 2 to 6. The samples were analyzed using the AbsoluteIDQ p150 kit (BIOCRATES Life Sciences AG, Austria). All the measurements were performed on a QTRAP 5500 (AB SCIEX, USA; flow injection analysis, ESI in both + and − MRM mode), and the data were processed in MetIQ software (AbsoluteIDQ kit). In total, 163 metabolites were quantified.

**MTBL59:** This data set is described in Franceschi et al and Wehrens et al and can be downloaded from the MetaboLight web page https://www.ebi.ac.uk/metabolights/MTBLS59. It contains 20 apple samples, which were analyzed by liquid chromatography–mass spectrometry (LC-MS). The first 10 samples of apples were analyzed without any modification. The second 10 samples were spiked with naturally occurring substances in apples. In that way, two groups with five known biomarkers were created and can be analyzed. Data preprocessing was carried out as outlined in Wehrens et al. Only the first 9 minutes of the chromatography was subtracted, leading to 197 features.

**PKU:** This data set concerns plasma samples from PKU patients ($n_1 = 27$) and healthy controls ($n_2 = 17$), where untargeted metabolomics analysis based on the work of Wang et al was performed in the Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacky University Olomouc, Czech Republic. The data were processed using the vendor software Compound Discoverer 3.0 (Thermo Fisher Scientific) and exported to the R software to perform correlation analysis to merge redundant features (RT difference $\leq 0.02$ min, $r \geq 0.95$, centered log ratio [clr] transformation) and to perform other statistical evaluation (the data were corrected using QC samples and LOESS regression; potential metabolites with CV higher than 30% were excluded from further data processing). There are in total 2336 features in the data set. On the basis of the biological analysis of the PKU disease, there are four known biomarkers.

3.2 | Visualization of cellwise outliers

We use the data set IMD to visualize cellwise outliers. In this multigroup data set, centering and scaling were performed according to the majority group, which is the control group; see Equation (7). Then, the cell-rPLR method was applied as described in Section 2. The resulting heatmap is shown in Figure 3 using the adjusted Tukey biweight function (8) and in Figure 4, which is based on the adjusted Hampel function (12). For reasons of space, we omitted the first 36 control patients from the visualization and showed only controls 37 to 50 and the patients from the different disease groups, separated by black horizontal lines. At a first glance, Figure 4 seems to provide a much clearer picture concerning potential cellwise outliers. This is due to the fact that the adjusted Hampel function assigns 0 to a much broader range of “normal” values than the adjusted Tukey biweight function. A red value corresponds to a “positive outlier” and to a value that is higher than expected. A blue value indicates a “negative outlier,” with a value lower than expected. Figure 3 shows that control observation 48 has systematic bias from the others: The first block of variables for this
sample is outlying in the positive direction, whereas the last block is outlying in the negative direction. This is not so clearly visible in Figure 4. If an entire signal of a sample is outlying, the sample preparation could have been done a bit different from the other samples, a different device setting could have been used during the analysis, or the control might have been biased by a certain unknown nonphysiological state. Moreover, variable PC.aa.C40.3 shows several negative outliers. However, the most visible (positive) outliers are in vertical blocks, indicated by the black rectangulars. In fact, these rectangular regions are the known metabolites C3 for group MMA and PA, Met for HCYS, and Phe for PKU. The first cell (patient MMA 34) of the biomarker C3 was not identified as outlier, which might mean that this patient is in an early stage of his/her disease or is possibly already cured. Note that positive outliers (red color) mean that the corresponding variables have increased values (dominance) for these observations.

This example shows that the choice of the outlyingness function should be based on the analysis goal. If a clear indication of outlyingness is desired, the adjusted Hampel or Huber function could be used, where Hampel would in general lead to more saturated colors. The adjusted Tukey biweight function would give more light red/blue color instead of white cells but indicate clearer small deviations from the “normal” behavior.

3.3 | Simulation study for outlier identification

In this section, we will more thoroughly test the algorithm cell-rPLR for outlier detection. This will be based on the data sets MTBL59 and PKU, where (additional) cellwise outliers will be generated by simulations, using

![FIGURE 3](image1)

**FIGURE 3** Outlier diagnostics for the IMD data, using the adjusted Tukey biweight function

![FIGURE 4](image2)

**FIGURE 4** Outlier diagnostics for the IMD data, using the adjusted Hampel function
\[ \tilde{x}_{ij} = x_{ij} \cdot M + A, \]  
(17)

where a multiplicative effect $M$ is generated from a uniform distribution $U(0.2, 0.5)$ or $U(2, 10)$ (with equal probability) and an additive effect $A$ is generated from $U(2, 5)$. This modification is done for a random percentage between 5% and 15% of the cells, which are randomly picked.

The modified cells are treated as “true” outliers, which should be identified with the algorithm. Note that by accident, scheme (17) could produce values that are still not very extreme and thus hard to identify as outliers. On the other hand, the data sets could already include cellwise outliers, but since they are unknown, an identification by the algorithm would count as a wrong decision.

We will compare our algorithm with the method detect deviating cells (DDC),\(^7\) which can be considered as a state-of-the-art cellwise outlier identification method. In the first step, DDC robustly standardize the columns of the data matrix, univariate outlier detection is applied to all variables separately, and outlying cells are marked. Later, the correlation structure is computed based on the nonmarked observations. This is followed by the prediction of each nonmarked data cell in the same row, considering only the correlating variables. The final outlyingness level is determined by the difference between the predicted and the reported values for each cell. The bigger the difference, the more outlying is the cell. In order to apply DDC appropriately, we first preprocess the data with the PQN transformation. The original version of DDC algorithm does not use the data label, and this version is also used in our simulations. However, the DDC could be redesigned in order to incorporate the group information.

In each of the 100 iterations, the receiver operating characteristic (ROC) curve was computed. The ROC curve shows the proportions of correctly identified outliers (sensitivity) and incorrectly identified nonoutliers (one minus specificity) for varying outlier cutoff points. Figure 5 shows the 100 different ROC curves for the cell-rPLR (with the adjusted Tukey biweight function) and the DDC algorithm for the data sets (a) MTBL59 and (b) PKU. A good method would lead to an ROC curve, which is close to the upper left corner of the plot (all outliers correctly identified, no false outlier indication). The plots reveal that the performance of the algorithms is better for the MTBL59 data set (197 variables) than for the PKU data set with much more variables (2336). In both data sets, the cell-rPLR algorithm clearly shows a better performance compared with the DDC method. Note, however, that the DDC method does not make use of the grouping information.

4 | PERFORMANCE OF CELL-RPLR FOR BIOMARKER IDENTIFICATION

As in the previous Section 3.3, we use the data sets MTBL59 and PKU to test the performance of the method cell-rPLR for identifying biomarkers. Note that these data sets only consist of two groups of observations, which is the setting we require for cell-rPLR biomarker identification. In particular, we are interested in the behavior of the method in the presence of contamination by cellwise outliers. Therefore, the same kind of data contamination is applied as introduced in (17), and for 100 simulation runs, 5%, 10%, 15%, 20%, and 25% of randomly selected cells of the data matrices are contaminated.
We will compare the following methods:

**cell-rPLR:** This algorithm is applied as described in Section 2.5. A ranked list for the variable importance is obtained by the comparison of the medians of the outlyingness values; see Equation (16). For a list of identified biomarkers, the permutation test as described in Section 2.5 is applied.

**PLS-VIP:** Partial least squares discriminant analysis (PLS-DA)\(^{30-32}\) is applied to the data set, which is doing PLS regression on the binary response containing the group labels. This method results in a projection of the samples on few latent variables. Since PLS-DA does not return the information of the most important variables for group discrimination, a variable importance projection (VIP) score\(^{33,34}\) is computed, which sums up the contribution of each variable in the model. The VIP scores are then sorted in ascending order so that a ranking of the variables is created. It is generally accepted that a variable should be marked as a biomarker if its VIP score is bigger than 1.\(^{35,36}\)

**PLS-SR:** There are many methods available for extracting the importance of the variables in a PLS-DA model.\(^{37}\) Here, we consider the selectivity ratio (SR),\(^{38,39}\) which is using scores and loadings from PLS-DA and computes a proportion of explained variance for each variable. Again, this results in a list of variables, sorted according to their importance, as well as in a list of identified biomarkers, which are selected as variables with SR score above the 0.95 quantile of the distribution \(F_{n-2,n-3}\).\(^ {40}\)

**PRM-VIP:** Partial least squares discriminant analysis is not robust against data outliers,\(^ {41}\) and thus, the robust counterpart based on partial robust M-regression (PRM)\(^ {42}\) is used. This is followed by computing the VIP score as a measure for variable importance.

**PRM-SR:** As before, PRM is applied but followed by computing the SR.

**DesEq:** This method builds on two steps: (a) an internal normalization of the variables by their geometric means and (b) a decision about the importance of variables.\(^ {43}\) For this purpose, a negative binomial generalized linear model is fit to each variable, and the \(P\) value from a Wald test\(^ {44,45}\) is computed for creating a rank

---

**FIGURE 6** A-D, Average ranks of the methods for the identification of the four known biomarkers in the PKU data, in a simulation setting with increasing amount of contamination
of importance for the variables and a list of identified biomarkers. The method removes outliers based on Cook’s distance.

Aldex: This method is based on Monte Carlo simulations of the Dirichlet multinomial model.\textsuperscript{46,47} The clr transformation\textsuperscript{48} is internally used. Then, $P$ values obtained from Welsch’s test\textsuperscript{49} are employed for ranking the variables and for returning a list of identified biomarkers.

For the methods employing PLS or PRM, a preprocessing step is necessary, and we decided to apply the PQN transformation,\textsuperscript{13} which is widely used.

Figures 6 and 7 show the results of the simulation study. The subplots in these figures report the resulting ranks for one particular known biomarker. Figure 6 refers to the results for the PKU data with four known biomarkers, and Figure 7 to those from the MTBL59 data with five known biomarkers. We report here the average ranks (over the 100 simulations) for the individual methods. Because of the large number of variables, the average ranks are shown on a base-10 logarithmic scale.

**FIGURE 7** A-E, Average ranks of the methods for the identification of the five known biomarkers in the MTBL59 data, in a simulation setting with increasing amount of contamination
Figures 6 and 7 show that for many of the methods, the average ranks for the known biomarkers increase with increasing data contamination. An exception is the cell-rPLR algorithm, which leads in general to the lowest ranks (at least for the contaminated situation) and also to stability in case of contamination. In particular for the PKU data (Figure 6), the PRM-based methods show a very poor performance, even for the uncontaminated data. This is because in this very high-dimensional data set, existing cellwise outliers might affect several observations and PRM then downweights all these observations. Especially when adding cellwise outliers, PRM leads to poor results exactly because of this reason; see also Figure 7. Depending on the specific biomarker, Aldex and DesEq also lead to reasonable performance, but they are clearly affected by the contamination. For the PKU data (Figure 6), PLS-SR is also quite competitive, but it completely fails for the PKU data.

Each of the considered methods returns the information if a variable is identified as a biomarker or not. In case of cell-rPLR, the permutation test (see Section 2) is employed to deliver this information. Thus, we evaluate the performance of correct biomarker identification based on the same simulation scenario as used before, for the data sets PKU and MTBL59. For each method, the true positive rate (TPR) as the proportion of correctly identified biomarkers (Sensitivity) and the true negative rate (TNR) as the proportion of correctly identified nonbiomarkers (Specificity) are computed. In the ideal case, both the TPR and TNR should yield values close to 1. Figure 8 summarizes the average values for TPR and TNR over the 100 simulation runs and for the MTBL59 data set. According to the false positive rate (FPR), the methods PLS-SR, Aldex, PRM-SR, and cell-rPLR show excellent behavior. However, PLS-SR has a very poor TPR (true biomarkers not identified). Also, the TPR of PRM-SR suffers from the contamination, because PRM can only cope with rowwise contamination. Aldex and DesEq are also sensitive to contamination, even the algorithm cell-rPLR shows a slight deterioration with increasing contamination—probably due to some effect of the permutation test. Overall, however, cell-rPLR is the clear winner under contamination but shows also competitive performance without contamination.

5 | SUMMARY AND CONCLUSIONS

This paper introduced a method called cell-rPLR, which allows to detect outlying cells in a data matrix of metabolomic data and also identifies biomarkers, even in presence of such cellwise outliers. This method does not require the usual data preprocessing (normalization and scaling) needed by many other methods, because the elemental information is pairwise log ratios between the variable values. The size effect that is often present in metabolomic data thus is automatically filtered out by using the log ratios. To the best of our knowledge, this is the first paper focusing on cellwise outliers in the context of metabolomic data. Cell-rPLR applies an outlyingness function to the robustly centered and scaled pairwise log ratios, and the results are outlyingness values for each observation and for each pair of variables (three-way array). After robust aggregation, one obtains an outlyingness values matrix of the same dimension as the data matrix, containing the cellwise outlyingness information. The outlyingness values are in the interval $[-1,1]$; outlyingness values around 0 point at “normal” data cells, and outlyingness values close to +1 or −1 indicate atypical data cells. This information can be visualized in a heatmap by using a color coding of the outlyingness values. The appearance of the heatmap depends a lot on the chosen outlyingness function, but the choice of the outlyingness
function would not essentially change a resulting ROC curve (this is valid at least for the outlyingness functions proposed in this paper). Thus, the heatmap can be regarded as a diagnostics tool for investigating the data structure. Cellwise outliers could indicate data problems, but they also indicate biomarkers if they are systematically present in one variable of a particular data group. A permutation test for biomarker identification was developed based on the cell-rPLR algorithm. On the basis of the simulations using artificially contaminated real data, the performance of cell-rPLR was compared with a state-of-the-art cellwise outlier detection algorithm, where it turned out that cell-rPLR was more accurate and at the same time it did spot fewer “normal” data cells as outliers. Similarly, simulations have shown that the performance of cell-rPLR in biomarker identification was competitive to alternative methods and superior in the presence of cellwise contamination.

The cell-rPLR is implemented in the software environment R\(^5^0\) and can be downloaded as package cellrPLR from https://github.com/walachja/cellrPLR. The package contains the implementation of the cell-rPLR algorithm and the permutation test for biomarker identification. Heatmaps for visualizing cellwise outliers are included as well. Furthermore, for an easier exploration and understanding of the data, a Shiny app is a part of the package. Shiny,\(^5^1\) an open source R package, is a web application and serves as an interactive tool for visualization. The Shiny app allows to interactively apply different outlyingness and aggregation functions for the cell-rPLR algorithm and supports zooming into regions of the data matrix to see more details. Moreover, the variables can be interactively ordered based on their importance for the group discrimination.

In our future work, we will focus on an automatic classification of cellwise outliers into technical artifacts (cellwise outliers) and biological artifacts (biomarkers) using the cell-rPLR algorithm. This is particularly important in case of very small groups of data, where this distinction is difficult with the current version of the algorithm. Moreover, the cell-rPLR method will be extended to biomarker identification for the multigroup case.

**ACKNOWLEDGMENTS**

The authors are grateful to anonymous referees for their constructive comments on an earlier version of the paper. This research is supported by the FWF (Austrian Science Fund) and the GACR (Czech Science Foundation) project number I1910-N26 (15-34613L) and GACR project number 18-12204S.

**ORCID**

Jan Walach https://orcid.org/0000-0002-7786-0493  
Peter Filzmoser https://orcid.org/0000-0002-8014-4682  
Štěpán Kouřil https://orcid.org/0000-0002-7964-1147  
Tomáš Adam https://orcid.org/0000-0003-4317-4708

**REFERENCES**

1. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS*. 2010;5(6):463.
2. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *JNCI: J Natl Cancer Inst*. 2001;93(14):1054-1061.
3. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*. 2009;26(3):392-398.
4. Huber PJ, Ronchetti EM. *Robust Statistics, Series in Probability and Mathematical Statistics*. New York, NY, USA: John Wiley; 1981.
5. Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M. *Robust Statistics: Theory and Methods (With R)*. Chichester, UK: Wiley; 2019.
6. Maronna R, Martin RD, Yohai V. *Robust Statistics*. Chichester, UK: John Wiley & Sons; 2006.
7. Rousseeuw PJ, Bossche WVD. Detecting deviating data cells. *Technometrics*. 2018;60(2):135-145.
8. Öllerer V, Alfons A, Croux C. The shooting S-estimator for robust regression. *Comput Stat*. 2016;31(3):829-844.
9. Warrack BM, Hnatyshyn S, Ott KH, et al. Normalization strategies for metabolomic analysis of urine samples. *J Chromatogr B*. 2009;877(5-6):547-552.
10. Filzmoser P, Walczak B. What can go wrong at the data normalization step for identification of biomarkers? *J Chromatogr A*. 2014;1362:194-205.
11. Kvalheim OM, Brakstad F, Liang Y. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Anal Chem*. 1994;66(1):43-51.
12. Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC. Scaling and normalization effects in NMR spectroscopic metabolomic data sets. Anal Chem. 2006;78(7):2262-2267.

13. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabolomics. Anal Chem. 2006;78(13):4281-4290.

14. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Modeling and Analysis of Compositional Data. Chichester, UK: John Wiley & Sons; 2015.

15. Walach J, Filizmoser P, Hron K. Data normalization and scaling: Consequences for the analysis in omics science. In: Jaumot J, Bedia C, Tauler R, eds. Data Analysis for Omics Sciences: Methods and Applications Amsterdam. The Netherlands: Elsevier; 2018: 65-196.

16. Beaton AE, Tukey JW. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics. 1974;16(2):147-185.

17. Yohai VJ, Zamar RH. High breakdown-point estimates of regression by means of the minimization of an efficient scale. J Am Stat Assoc. 1988;83(402):406-413.

18. Maronna RA, Zamar RH. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics. 2002;44(4):307-317.

19. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust Statistics: The Approach Based on Influence Functions. New York, NY, USA: John Wiley & Sons; 1986.

20. Fisher RA. The Design of Experiments. UK: Oliver & Boyd, Edinburgh and London; 1935.

21. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. J Am Stat Assoc. 1980;75(371):591-593.

22. Janečková H, Hron K, Wojtowicz P, et al. Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders. J Chromatogr A. 2012;1226:11-17.

23. Franceschi P, Masuero D, Vrhovsek U, Mattivi F, Wehrens R. A benchmark spike-in data set for biomarker identification in metabolomics. J Chemom. 2012;26(1-2):16-24.

24. Wehrens R, Franceschi P, Vrhovsek U, Mattivi F. Stability-based biomarker selection. Anal Chim Acta. 2011;705(1-2):15-23.

25. Wang J, Christison TT, Misuno K, et al. Metabolomic profiling of anionic metabolites in head and neck cancer cells by capillary ion chromatography with orbitrap mass spectrometry. Anal Chem. 2014;86(10):5116-5124.

26. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc. 1988;83(403):596-610.

27. Miller MJ, Kennedy AD, Eckhart AD, et al. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. J Inherit Metab Dis. 2015;38(6):1029-1039.

28. Jansen RS, Addie R, Merkx R, et al. N-lactoyl-amino acids are ubiquitous metabolites that originate from CNDP2-mediated reverse proteolysis of lactate and amino acids. Proc Natl Acad Sci. 2015;112(21):6601-6606.

29. Václavík J, Coene KL, Vrobel I, et al. Structural elucidation of novel biomarkers of known metabolic disorders based on multistage fragmentation mass spectra. J Inherit Metab Dis. 2018;41(3):407-414.

30. Vold H. Path models with latent variables: the NIPALS approach. In: Blalock HM, Aganbegian A, Borodkin FM, Boudon R, Capecchi V, eds. Quantitative Sociology International Perspectives on Mathematical and Statistical Modeling. Academic Press: London, UK; 1975:307-357.

31. Wold S, Martens H, Wold H. Partial least squares projections to latent structures. 3D QSAR. 1993;1:523.

32. Stähle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. J Chemom. 1987;1(3):185-196.

33. Favilla S, Durante C, Vigni ML, Cocchi M. Assessing feature relevance in NPLS models by VIP. Chemom Intell Lab Syst. 2013;129:76-86.

34. Wold S, Johansson E, Cocchi M. PLS—partial least squares projections to latent structures. 3D QSAR. Drug Des. 1993;1:523-550.

35. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Syst. 2005;78(1):103-112.

36. Gosselin Ryan, Rodrigue Denis, Duchesne Carl. A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. Chemom Intell Lab Syst. 2010;100(1):12-21.

37. Mehmood T, Liland KH, Snipen L, Sebo S. A review of variable selection methods in partial least squares regression. Chemom Intell Lab Syst. 2012;118:62-69.

38. Rajalahlé T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemom Intell Lab Syst. 2009;95(1):35-48.

39. Kvalheim OM. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. J Chemom. 2009;24(7-8):496-504.

40. Rajalahlé T, Arneberg R, Kroksveen AC, Berle M, Myhr KM, Kvalheim OM. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. Anal Chem. 2009;81(7):2581-2590.
41. Filzmoser P, Se neels S, Maronna R, Van Espen PJ. Robust multivariate methods in chemometrics. In: Walczak B, Ferre RT, Brown S, eds. Comprehensive Chemometrics (vol. 3). Oxford, UK: Oxford, UK; 2009:681-722.
42. Serneels S, Croux C, Filzmoser P, Van Espen PJ. Partial robust M-regression. Chemom Intell Lab Syst. 2005;79(1-2):55-64.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
44. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc. 1943;54(3):426-482.
45. Harrell FE. Regression Modeling Strategies. Germany: Springer, Cham; 2014.
46. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS ONE. 2013;8(7):e67019.
47. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016;62(8):692-703.
48. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B Methodol. 1982;44(2):139-177.
49. Welch BL. The significance of the difference between two means when the population variances are unequal. Biometrika. 1938;29(3/4):350-362.
50. R Core Team. R: a language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/; 2018.
51. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. Shiny: web application framework for R. https://CRAN.Rproject.org/package=shiny, r package version1.1.0.; 2018.

How to cite this article: Walach J, Filzmoser P, Kouřil S., Friedecký D, Adam T. Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios. Journal of Chemometrics. 2020;34:e3182. https://doi.org/10.1002/cem.3182