TECHNISCHE
UNIVERSITÄT
WIEN
Vienna | Austria

Robust Log-ratio Methods for Classifying High-dimensional Metabolomics Data

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Univ.-Prof. Dipl.-Ing. Dr. techn. Peter Filzmoser, Institut für Stochastik und Wirtschaftsmathematik (E105)

eingereicht an der Technischen Universität Wien an der Fakultät für Mathematik and Geoinformation

von

Mgr. Jan Walach
Matrikelnummer 01528856

Diese Dissertation haben begutachtet:

doc. RNDr. Eva Fišerová, Ph.D.	dhr. dr. Johan A. Westerhuis

Wien, 15. Jänner 2019

Mgr. Jan Walach

Erklärung zur Verfassung der Arbeit

Mgr. Jan Walach
Untere Viaduktgasse 21,1030 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. Jänner 2019

Mgr. Jan Walach

Acknowledgements

I would like to thank my advisor Professor Peter Filzmoser for his guidance and support through the last years of my studies. Not only your academic help made my life so much easier. I can not thank you enough for enabling me to study in Vienna, where I could spent three and a half amazing years.

My many thanks go to my present and former colleagues for those fruitful discussions, advice and all the laughs we had together.

I could never be who I am today without my loving parents Zdeněk and Dana, to whom I owe so much. Last, but definitely not least, I will always be indebted to Klárka, mom of my son Ondra and my soulmate. You have always given me motivation and strength at the bottoms and cheered with me on the tops.

Abstract

The development of statistical methods which are able to deal with high-dimensional data belongs to the major research activities in statistics. In many fields (e.g. chemometrics, genomics, metabolomics) it is easy to measure and store data by using advanced modern techniques. Thus, there are also numerous real-world applications justifying these developments. One possible way how to deal with such data comes from the log-ratio point of view. There is whole branch of statistics devoted to log-ratios – Compositional Data Analysis. Compositional data represent a special type of multivariate data which describe parts of a whole. In this context only relative information is important. Because of these special features of compositional data, the application of standard statistical methods could lead to invalid conclusions.

The primary aim of the thesis is to introduce procedures for analysing high-dimensional data which originate from different groups. The main focus is set on applications in the field of metabolomics, where the different data groups consist of observations related to different diseases. The new methods should not only allow to differentiate between the groups, but they should also enable feature selection: only those features (variables), which allow to discriminate between the different groups, should be identified. An important request for these methods is their robustness against outlying observations, which is a common situation in real data.

Another interest of the thesis is the investigation of outliers in the data. We focus on both observational outliers and on so-called cell outliers. The former refers to the situation when an observation deviates from the majority of a group in possibly all variables, while in the latter case for a certain observation only the values in some variables (cells) are deviating. This will contribute to gain a better insight into the data structure.

Kurzfassung

Die Entwicklung von statistischen Methoden für hochdimensionale Daten gehört heute zu den wichtigen Forschungsaktivitäten in der Statistik. In vielen Disziplinen (z.B. Chemometrie, Genomik, Metabolomik) ist es einfach geworden, mit fortgeschrittenen modernen Techniken Daten zu messen und zu speichern. Daher gibt es auch unzählige Anwendungen in der Praxis, die solche Entwicklungen rechtfertigen. Eine Möglichkeit, solche Daten handhaben zu können kommt von der Sichtweise von logarithmischen Verhältnissen (*log-ratios*). Es gibt einen ganzen Bereich in der Statistik, der sich *log-ratios* widmet – die Analyse von Kompositionsdaten. Solche Daten sind spezielle multivariate Daten, die Anteile an einem Ganzen beschreiben. In diesem Zusammenhang ist nur die relative Information wichtig. Aufgrund der spezifischen Charakteristik von Kompositionsdaten könnte die Anwendung von herkömmlichen statistischen Methoden zu falschen Schlussfolgerungen führen.

Primäres Ziel der Dissertationsschrift ist es, Verfahren für die Analyse hochdimensionaler Daten, die von verschiedenen Gruppen kommen können, zu entwickeln. Hauptaugenmerk ist auf Anwendungen aus dem Gebiet der Metabolomik gelegt, wo die unterschiedlichen Datengruppen aus Beobachtungen bestehen, die zu unterschiedlichen Erkrankungen in Bezug stehen. Die neuen Methoden sollten es nicht nur ermöglichen, die Gruppen unterscheiden zu können, sie sollten auch eine Variablenselektion ermöglichen. Nur jene Merkmale (Variablen), die eine Unterscheidung der Gruppen ermöglichen, sollten identifiziert werden. Eine wichtige Anforderung an diese Methoden ist die Robustheit gegenüber Ausreißern, welche in Echtdaten häufig vorkommen.

Ein weiterer Schwerpunkt dieser Arbeit ist die Untersuchung von Ausreißern in den Daten. Der Fokus liegt sowohl auf Ausreißern in Form von ganzen Beobachtungen, als auch auf sogenannten zellweisen Ausreißern. Erstere Art von Ausreißern bezieht sich auf den Fall, in dem eine Beobachtung von der Mehrheit einer Gruppe möglicherweise in allen Variablen abweicht. Im letzteren Fall weicht die Beobachtung nur in bestimmten Variablen ab. Dies trägt dazu bei, einen besseren Einblick in die Datenstruktur zu erhalten.

Contents

CHAPTER 1

Introduction

There are several main tasks in the field of statistics and data science. Firstly, *clustering* corresponds with the analysis of data, where the response variable (the so called label) is not known or at least not used for the analysis. Secondly, *regression* tries to model a quantitative response based on a given set of predictors. Thirdly, *classification* deals with the same issue as regression with the difference that the response variable acquires only several discrete values. Fourthly, *outlier identification* analyzes the data in order to find observations which, due to various reasons, deviate from the majority .

The thesis focuses on the last two of the four situations mentioned above – *classification* and *outlier identification* and applies these methods mainly to data coming from the field of Metabolomics. Metabolomics aims at a systematic study of metabolites, their interactions, changes and responses to different kinds of stimuli – medicament, diets or drugs. Data coming from metabolomics, as well as from many other fields and applications, have a large number of variables which might lead to a failure of many classification and outlier identification methods. Most of the times, the number of analyzed samples is much smaller than the number of measured variables ($n \ll d$). Such a situation is referred to as flat data. The series of problems connected with an increasing number of measured variables is well examined and is generally known as *course of dimensionality*, see Beyer et al. (1999); Bennett et al. (1999). With increasing number of variables, the data space gets emptier and emptier and one would need an extensive set of samples to describe the space sufficiently. It might not be feasible due to costs or other limits. It is also easy to overfit the statistical model because of randomness effects, which are added with each additional non-informative variable. Furthermore, the vast majority of variables might be irrelevant and thus redundant in high-dimensional settings. This

creates a new problem of an appropriate identification of important variables – *feature selection*. It is called *biomarker identification* in metabolomics, and it is the third and the main topic of the thesis.

The goal of this chapter is to demonstrate readers several problems connected to the analysis of data mainly from metabolomics as well as to provide an overview of existing methods, their advantages and disadvantages.

1.1 Metabolomics

Metabolomics is a field of biochemistry. Together with genomics, transcriptomics or proteomics it is a part of the "omics" family. Metabolomics aims at a systematic study of metabolites, their interactions, changes and responses to different kinds of stimuli. It analyses a variety of living organisms from humans to bacteria and plants. All types of biological samples from bifluids, such as urine, blood or plasma, can be used for the analysis. Metabolomics is associated with metaboloms and metabolites. A metabolite is defined as a molecular mass organic compound, intermediate or metabolism product found in an organism (Oliver et al., 1998). Basically, they are small molecules which are being transformed during metabolism processes. A metabolome is a complete set of metabolites in the organism. Metabolomics analyzes a metabolome under a given condition, where the condition simply refers to the state of the organism (e.g. sick or healthy patient, tea before and after fermentation, etc.). Unlike genes or proteins, metabolites are directly related to biochemical activity. Hence, they can be more easily assigned to a certain condition. Therefore, metabolomics is a powerful tool of clinical diagnostics (Patti et al., 2012). Increasing popularity and importance of metabolomics is shown in Figure 1.1. The figure plots the growing number of publications found at the PubMed webpage for the keywords "metabolomics" and "biomarker" between the years 2002 and 2018.

Metabolomics can be divided into the *targeted* and *untargeted* approach (Patti et al., 2012). The choice of the approach strongly depends on the experimental objective and on available resources. Targeted analysis is an approach where specific metabolites are measured. It is usually driven by a concrete biochemical question. The number of metabolites is typically an order of magnitude lower (in hundreds) compared to the untargeted approach (in thousands). Hence, the interpretation of targeted analysis data is generally an easier task (Wu et al., 2011; Dudley et al., 2010). It is used e.g. for pharmacokinetic studies of drug metabolism or generic modification on a certain enzyme Nicholson et al. (2002). On the other hand, the costs of targeted analysis is

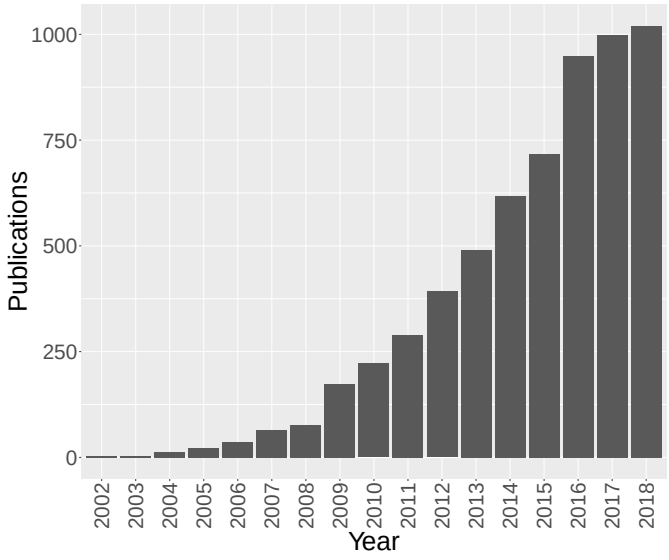


Figure 1.1: PubMed search using keywords "metabolomics" and "biomarkers", years 2002 to 2018. (Data collected on *14.1.2019*)

higher due to the fact that metabolites need to be adjusted to the machine before the analysis. It is done by commercially available chemical standards.

The untargeted approach analyzes metabolites which are not known beforehand. Instead of directly identifying them, chromatograhic peaks are found. Chromatograhic peaks are potential metabolites. However, these peaks must be firstly identified, which is a complicated process. For more details see Alonso et al. (2015); Gardlo (2016). Nonetheless, the advantage of untargeted analysis is that it identifies almost all metabolites. It makes it an ideal tool to provide an insight into complicated biological processes due to higher complexity of the data.

1.2 Data in metabolomics

Data from metabolomics have a typical structure with n observations stored in the rows and d variables in the columns of a data matrix. Thus, the data matrix $\mathbf{X}$ consists of elements x_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Thanks to powerful analytical tools and relatively high costs, the number of variables is usually much higher than the number of observations. In a typical metabolomic setting, the samples in $\mathbf{X}$ are coming from $G = 2$ groups. For example, the first group might refer to healthy people, and the second group contains patients with a certain condition.

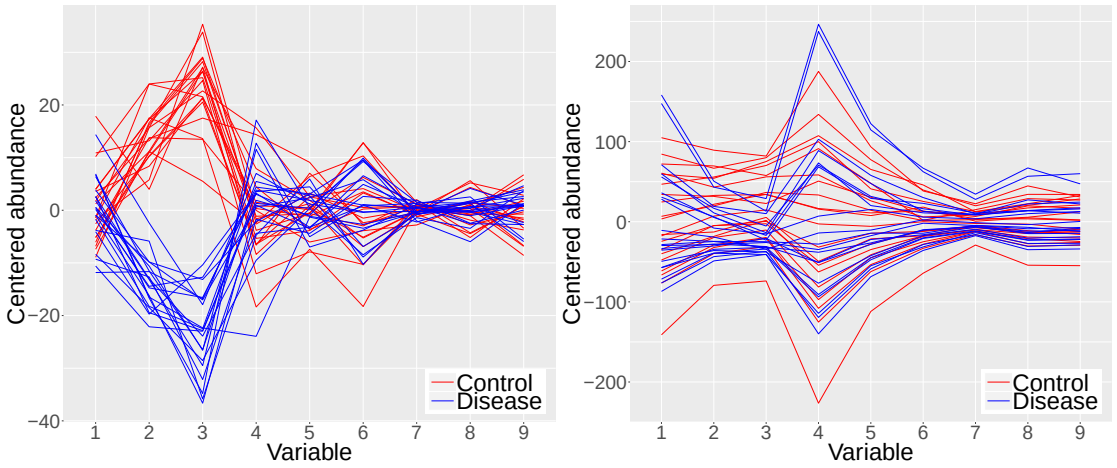
The most important information in the dataset is usually contained in its variability.

For that reason it is important to describe and understand different types of variability in metabolomics. Biological variability describes differences within the subjects. Each sample is up to some point unique. Thus, if, for example, two samples of tissue are analyzed, the resulting abundances will slightly differ. Secondly, technical variation is caused either by analytical errors or errors and differences concerning the machine used for the analysis itself. In metabolomics, two powerful techniques are used to perform the analysis: Nuclear Magnetic Resonance (NMR) imaging and different types of Mass Spectrometry (MS). MS techniques are the most sensitive for the simultaneous analysis of a large number of compounds. NMR complements mass spectrometry. Its sensitivity is lower, but it is uniquely capable of elucidating molecular structure (Burgess et al., 2014). These two tools produce big data in the form of abundances of a compound of a biological sample. The third type of variation is called induced biological variation. Two samples in different state (e.g. control versus disease) will differ in some of the measured entities exactly due to its diverse states. This type of variation is of great interest. Finding, examining and understanding these differences is one of the main goals of metabolomics.

Many statistical methods are based on the assumption that errors are fluctuating around zero with certain constant variation. However, this is frequently not the case, since often increasing abundances correspond to increasing noise. Such a situation is referred to as *heteroskedasticity* and it needs to be addressed prior to the statistical analysis.

Data in metabolomics highly depend on the material (blood, tissue, urine, etc.) used for measurements. There are some issues connected with the type of the analyzed material. One of them is the so-called size–effect. It refers to a situation when the samples cannot be directly compared, because of differences in their volumes. A typical example is the analysis of urine or blood. Naturally, the urine concentration strongly depends on the level of water intake and on other physiological factors. Therefore, the abundances of metabolites in different urine samples differ. An analogous situation concerning blood samples is referred to as "dry blood spots". Similarly, one can imagine that the analyzed amount is crucial. If, for example, two units of a blood sample are measured, it is expected that the abundance levels will be approximately twice as high for all metabolites compared to a measurement of one unit. A simulated toy example based on Filzmoser and Walczak (2014) illustrating the size–effect in the data is shown in Figure 1.2. Figure 1.2(a) describes a situation where the size–effect is not present in the data. There are two groups of samples – controls and disease patients. Nine variables (metabolites) are plotted on the x-axis. The y-axis shows the abundance, which is furthermore centered by its mean. Variables 2 and 3 obviously discriminate between the

two groups, hence they are considered to be biomarkers. The rest of the variables have no discrimination power and represent noise. However, as demonstrated in Figure 1.2(b), as soon as the size–effect is present in the data, the differences between the groups can vanish. Such a situation is not desirable and needs to be dealt with.



(a) Without size–effect

(b) With size–effect

Figure 1.2: Simulated data example: centered abundances of nine variables in a situation with and without size–effect.

There are several possibilities how to eliminate size–effect, heteroskedasticity and other data-related issues. In practice, normalization, scaling or transformation methods are used. For simplicity, they are divided here into (1.) normalization to internal standard methods, (2.) data–driven methods and (3.) methods based on log-ratios. Typically, the log-ratio methodology is considered to be part of the data–driven methods, but since it is a crucial concept throughout the thesis it is treated as a separate group.

"Housekeeping variable" normalization

The first type of methods is also called normalization to a "housekeeping variable". A reference variable is chosen due to its stability throughout various samples. All the other variables are then divided by its reference intensity. Creatinine is a specific metabolite which appears in all urine samples. It is a chemical waste product, which is filtered by kidneys and is eliminated in urine. Under normal circumstances, creatinine is directly proportional to the urine concentration. However, the level of creatinine might be influenced by external factors, for instance kidney diseases (Garde et al., 2004; Warrack et al., 2009; Waikar et al., 2010), which might not be known prior to the analysis.

Consequently, such a normalization would result in biased conclusions.

Data–driven methods

Data–driven methods are a broad range of normalization, scaling and transformation approaches. A closer description is provided in Chapter 2. Here, the focus is on the frequently used Total Sum Normalization (TSN) (Craig et al., 2006; Giraudeau et al., 2014; Filzmoser and Walczak, 2014) and v (PQN) Dieterle et al. (2006). The TSN normalizes the data by forcing the sum of all variables of each sample to be constant (e.g. 1 or 100%). Afterwards, the transformed variables represent fractions or percentages of the whole. It is accomplished by dividing each abundance by the sum of the intensities of the whole sample. Thus, for each sample $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, $i = 1, \ldots, n$, TSN is constructed as

$$\mathbf{x}_i^{TSN} = \left(\frac{x_{i1}}{t_i}, \frac{x_{i2}}{t_i}, \ldots, \frac{x_{id}}{t_i} \right), \qquad \text{where } t_i = \sum_{j=1}^{d} x_{ij}. \tag{1.1}$$

The TSN can be biased and lead to unreliable results (Filzmoser and Walczak, 2014). This is the case especially in "far–from–closure" situations, where the sum of all intensities of each sample is generally different. In order to allow one or several measured intensities to increase in a constant sum scenario, others need to decrease, which might not necessarily correspond to the reality. Furthermore, the differences are redistributed equally among all other variables. It leads to creating artificial differences in variables and spoiling the feature selection analysis. The described problems are demonstrated in Figure 1.3(a). It follows the simulated toy example from Figure 1.2. The figure shows data after TSN. The discrimination power of the true biomarker variables 2 and 3 is visible again. However, the differences were indeed redistributed among all the other variables. Thus, one could falsely conclude that all variables contribute to discriminate between the groups, even thought in reality they are just noise variables.

Probabilistic Quotient Normalization (PQN) is a linear normalization method, which mainly aims to deal with the size–effect. The assumption of PQN is that the size–effect influences the whole sample and that biological changes between samples in measured concentrations affect only some parts of the spectra. As a practical consequence of this, PQN assumes that at most half of the variables are discriminating between the groups, and thus are biomarkers. If more than a half of the measured concentrations are biomarkers, PQN cannot be correctly applied.

PQN can be described as follows. Firstly, a reference spectrum is determined. In

practice, this is usually done by computing the median for each variable,

$$x_j^{ref} = \text{median}\left\{x_{1j}, \ldots, x_{nj}\right\},$$

for $j = 1, \ldots, d$. Secondly, the ratios of all variables with the corresponding reference spectrum are taken. Furthermore, the median of these values is computed for each sample, thus

$$x_i^* = \text{median}\left\{\frac{x_{i1}}{x_1^{ref}}, \ldots, \frac{x_{id}}{x_d^{ref}}\right\},$$

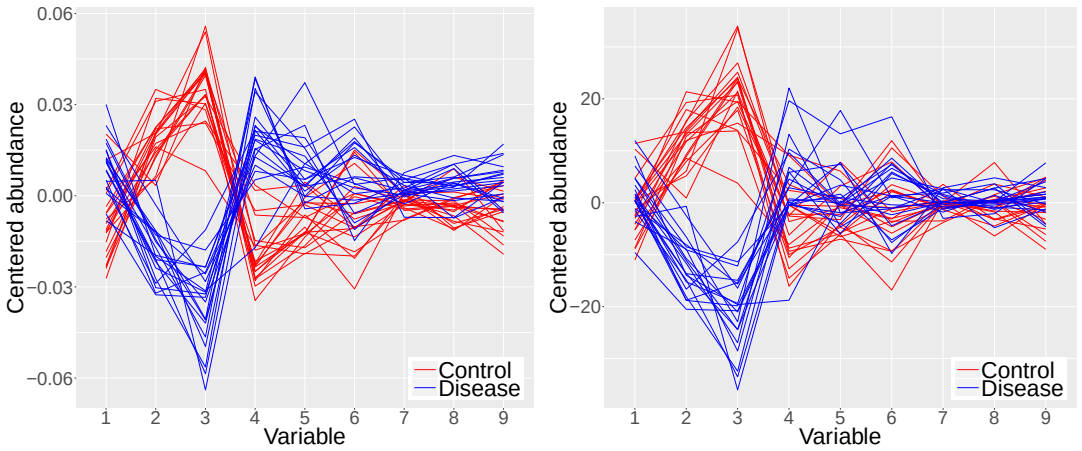
for $i = 1, \ldots, n$. The computed values are the estimation of the size–effect for each sample. Thus, the division by the estimation of the size–effect normalizes the samples to the same concentration level

$$x_{ij}^{PQN} = \frac{x_{ij}}{x_i^*} \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, d. \tag{1.2}$$

As demonstrated in Figure 1.3(b), PQN can indeed deal with the size–effect, since the differences are visible for the biomarker variables 2 and 3. Furthermore, the rest of the variables could be interpreted as noise variables, with the exception of variable number 4, which still shows some discrimination power. However, as shown in Chapter 2, PQN does not necessarily give the best performance in real-world metabolomic datasets. This is probably due to the strict PQN assumption that all measured variables have the same concentration level for each sample. The assumption arises from the last step of PQN (Equation (1.2)), where each sample is divided by its estimated level of concentration. However, in reality, the size–effect is more complicated because of the random noise which influences the concentration levels of each variable.

Methods based on log-ratios

The third group of normalization methods is based on log-ratios. The idea behind the log-ratio methodology is that the sum of parts of each sample does not carry any important information. Rather, relative information stored between the ratios contains relevant insights. Thus, the sum is only a representation, not an inherent property of the data. There is a whole part of statistics called *Compositional Data analysis* (CoDa), which focuses on data containing relative information, see, for example, Aitchison (1989); Pawlowsky-Glahn et al. (2015); Filzmoser et al. (2018). Such data are called compositional data or compositions. Compositions consist of strictly positive values which are part of a whole (Pawlowsky-Glahn et al., 2015). Percentages, frequencies or concentrations are a typical example of compositional data. If the compositional structure of data is not



(a) TSN normalization (b) PQN normalization

Figure 1.3: Simulated data example: centered abundances with size–effect after TSN and PQN transformation.

considered and compositions are treated as absolute values, as in a common approach, classical statistical methods can lead to biased results.

Ratios are scale (and size) invariant. However, due to the asymmetry of their variance, $\mathrm{var}\,(x_1/x_2) \neq \mathrm{var}\,(x_2/x_1)$, they are not easy to work with. This deficiency can be overcome if instead of ratios log-ratios are used. Then, $\mathrm{var}\,(\log\,(x_1/x_2))$ is equal to $\mathrm{var}\,(\log\,(x_2/x_1))$. The log-ratio methodology is the basis of CoDa. The geometry for compositional data is not the usual Euclidean one, but rather the Aitchison geometry on the simplex (Aitchison, 1989). Log-ratios take out the problem of a constrained sample space. Thus, closure of the data is not important. Note that log-ratios of original and of already normalized data (e.g. by total sum normalization) yield the same result.

Datasets from Metabolomics are not compositional in a sense that they have constant sum, since an increase of a certain metabolite does not imply a decrease of others. Nonetheless, the size–effect makes the absolute values irrelevant and the important information is indeed relative. Thus, metabolomics data can be viewed and treated as compositions. Furthermore, the size–effect can be removed by working with log-ratios, since $\log\left(\frac{s\cdot x_1}{s\cdot x_1}\right) = \log\left(\frac{x_1}{x_1}\right)$, where s represents the size–effect.

There are several popular CoDa transformations in compositional data analysis, including additive, isometric or centered log-ratio (Filzmoser et al., 2018). The centered log-ratio transformation is probably the most used CoDa transformation, due to its good performance and relatively easy interpretability. The clr transformation moves the data from the simplex to the usual Euclidean geometry (Filzmoser et al., 2018). The ith

observation $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ is transformed to

$$\mathbf{x}_i^{clr} = \left(x_{i1}^{clr}, \ldots, x_{id}^{clr}\right) = \left(\log\left(\frac{x_{i1}}{g(x_i)}\right), \ldots, \log\left(\frac{x_{id}}{g(x_i)}\right)\right), \tag{1.3}$$

where $g(\mathbf{x}_i) = \sqrt[d]{\prod_{j=1}^{d} x_{ij}}$ is the geometric mean of the ith observation, for $i = 1, \ldots, n$. Thus, clr transformed data have the same dimension d as the original dataset, however, the components sum up to zero, $x_{i1}^{clr} + \ldots + x_{id}^{clr} = 0$. This means that clr transformed data do not have full rank d, which could create problems for methods like discriminant analysis, where a covariance matrix with full rank is required. Also for some robust statistical methods this is a prerequisite. On the other hand, the components of clr transformed data have a straightforward interpretation in terms of a dominance of the corresponding compositional part on an average behavior (geometric mean) of the values in the composition. Note that clr transformed data cover log-ratio information of all different pairs of variables: for example, the first component can be written as

$$x_{i1}^{clr} = \frac{1}{d}\left(\log\left(\frac{x_{i1}}{x_{i2}}\right) + \ldots + \log\left(\frac{x_{i1}}{x_{id}}\right)\right).$$

Nonetheless, as demonstrated in Figure 1.4, the clr transformation might enhance differences between groups and create artificial biomarkers. At the same time, the within-group variance is decreased. The reason is that the geometric mean is used as a denominator in Equation 1.3 for transforming all variables, and the geometric mean is computed from abundances also from biomarker variables, which creates the differences. Although with an increasing number of non-biomarker variables in the data, the geometric mean is less and less affected, it is still influenced.

To solve this problem one can work with pairwise log-ratios. Log-ratios hold the scale invariance property. Furthermore, the size effect does not play any role if pairwise log-ratios are considered. The ith sample of the dataset is described by pairwise log-ratios, organized in a matrix $\mathbf{R}_i \in \mathbb{R}^{d \times d}$ as

$$\mathbf{R}_i = [r_{jk}] = \left[\log\left(\frac{x_{ij}}{x_{ik}}\right)\right], \tag{1.4}$$

where $j, k = 1, \ldots, d$. The diagonal elements are equal to 0 and $r_{jk} = -r_{kj}$, since $\log\left(\frac{x_{ij}}{x_{ik}}\right) = -\log\left(\frac{x_{ik}}{x_{ij}}\right)$. Thus it is sufficient to only consider the upper (or lower) triangular matrix of $\mathbf{R}_i$. This information can be unfolded into a vector of length $d \times (d-1)/2$, containing the pairwise log-ratio information of the $i$the observation. If this is done for every observation with index $i = 1, \ldots, n$, a matrix of dimension $n \times d \cdot (d-1)/2$ is resulting. Although the dimensionality increases substantially compared to the original data, usually only few biomarker variables are present, and they are reflected in only

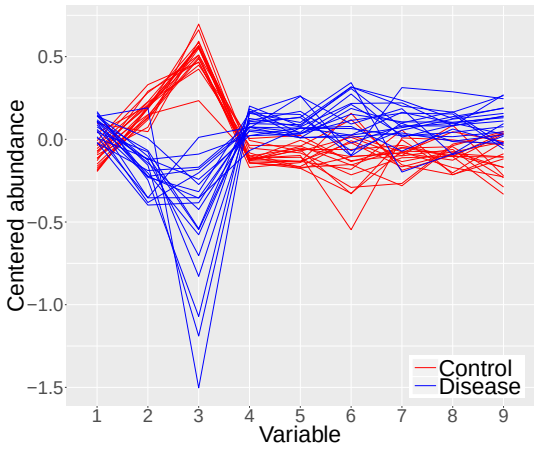


Figure 1.4: Simulated data example: centered abundances with size–effect after CLR transformation.

few log-ratios. All remaining log-ratios are not affected by the biomarkers. This is demonstrated in the simulated toy example in Figure 1.5. The x-axis shows all different pairwise log-ratios and the y-axis corresponds to the centered abundances. The differences between the two groups are visible only in log-ratios which contain either biomarker variable 2 or 3.

The disadvantage of methods based on log-ratios is that strictly positive values are required. After appropriate preprocessing, metabolomic data should not have negative values, however, zero values might be present in the data. Firstly, one should keep in mind that there are several reasons why there are zero values in the data. It might be simply because the measured value is not present at all. This zero type is called essential zero. However, in omics fields the zeros typically appear as values below detection limit, so-called rounded zeros, or they arise after a pre-treatment step: if the value is below a certain threshold, it is suppressed to zero because of the inaccuracy of the measurement device.

There are several methods available dealing with zero imputation. The easiest method is to replace all the zeros with 2/3 of the detection limit of the measurement device. However, if many zeros are present, they are all replaced by the same value, thus lowering the variability of the data set. There are also several CoDa methods (e.g. Templ et al. (2016); Martín-Fernández et al. (2012, 2015)) for the imputation of zero values, based e.g. on the covariance structure of the data. More details about the zero imputation strategies are given in Section 2.4.2 of the thesis.

The methods based on pairwise log-ratios are a crucial part of the thesis and are

more deeply described in Chapters 2, 3 and 4.

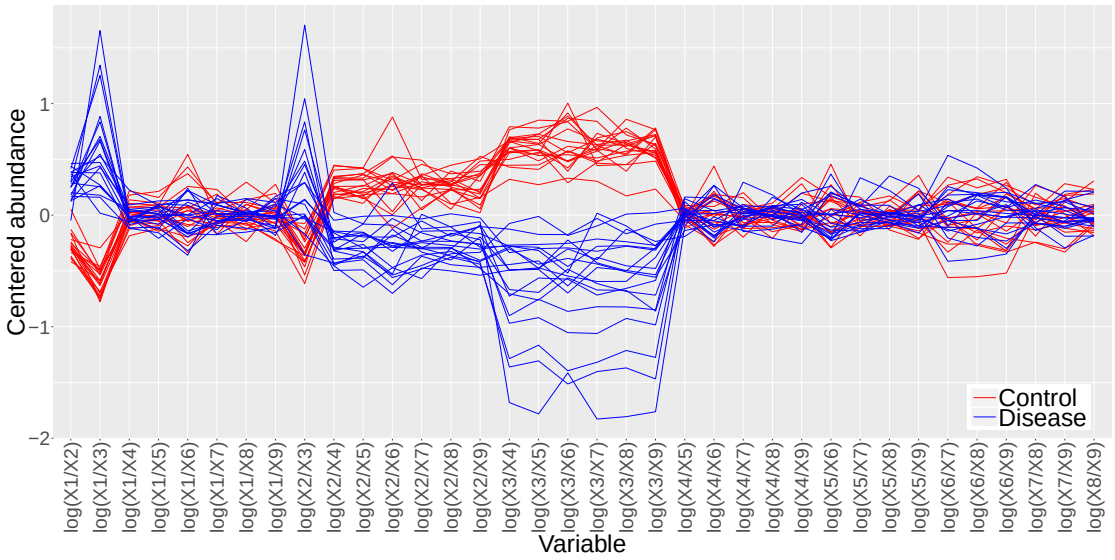


Figure 1.5: Simulated data example: centered abundances with size–effect, pairwise log-ratios were computed.

1.3 Feature selection methods

High-dimensional data often contain a large number or even vast majority of noise variables. In order to interpret a fitted statistical model correctly, it is often desired to identify only those variables which carry relevant information. In metabolomics, those variables are called biomarkers, and their identification is one of the most important tasks in this field, see, for example Roessner and Dias (2016). Biomarkers do have a biological reason for discriminating between the groups of e.g. healthy persons and diseased patients.

Feature selection methods can be categorized into feature subset selection and feature extraction methods (Hira and Gillies, 2015). Feature subset selection operates by removing variables which do not contribute or are redundant. Furthermore, they are separated into filters, wrappers, embedded approaches, see Guyon and Elisseeff (2003); Saeys et al. (2007). Feature extraction creates new features as a combination of original variables in order to lower the dimensionality of selected variables.

Feature subset selection

Filters evaluate variables based on internal properties of the data. Classically, some kind of score is calculated and only high-scoring variables are kept. Then, these features are an input to the classification method. An advantage of filters is that they are computationally simple and fast even for very high-dimensional datasets. Moreover, they do not depend on the classification method. Thus, feature selection is performed only once and then the selected subset of features can be analyzed by different classifiers. On the other hand, filters usually ignore interactions among variables. Hence, only univariate relationships are considered, which could result in worse performance. There are many filter methods including t-tests (Student, 1908), Information Gain (Yang et al., 2010) or the recently proposed ETC method (Schroeder, 2018) .

Wrapper methods evaluate subsets of variables based on their predictive performance. In practice, two things need to be applied. Firstly, a strategy how to search among all possible subsets of variables, and secondly, a strategy how to asses the prediction performance. The best subset selection is referred to as "brute-force" approach, since it computes all possible combinations of d variables. However, this includes the assessment of 2^d possible models which is computationally unbearable, even for a smaller number of variables. Considering only 30 variables, the number of possible combinations is $2^{30} \approx 10^9$. Such a problem is referred to as NP-hard (Amaldi and Kann, 1998). Nonetheless, other approaches such as backward or forward selection (Kittler, 1986) or generic algorithms (Holland, 1992) efficiently examine some subsets of variables. Thus, they could be used also for higher dimensional data. Assuming that prediction methods are used as a black box, an advantage of wrapper methods is that they are exceptionally universal and simple to use.

Variable selection of embedded approaches is inherent to the method. Thus, it is performed as a part of the training of the statistical method. They are more efficient than wrappers, since they avoid retraining of a predictor for every considered variable subset. Typically, an objective function of embedded methods consists of two parts – goodness-of-fit of the model and penalty term. The latter serve as a reduction of the number of variables used in the model. For example, the least absolute shrinkage and selection operator (LASSO) sets a penalty term to the problem of minimizing the residual sum-of-squares as a sum of absolute values of variable coefficients. This is forcing some of the coefficients to be equal to zero, and consequently they do not participate in the model, see Tibshirani (1996). Additional embedded approaches such as random forests have a build-in mechanism how to perform variable selection Guyon and Elisseeff (2003). The disadvantage of embedded methods is that they are more likely to overfit the number

of selected variables (Lal et al., 2006).

Feature extraction

Feature extraction creates a new, smaller set of variables, which captures most of the useful information in the data. It assumes that the data lie on a low-dimensional subspace. Most methods create so-called latent variables as a linear combination. However, there are also non-linear feature extraction methods such as self-organizing maps Kohonen (1982). Feature extraction methods can be unsupervised, such as the well-known PCA (Principal Component Analysis), or supervised, e.g. PLS (Partial Least Squares) regression.

To demonstrate how feature extraction methods work, the classical multivariate statistical tool PCA Jolliffe (2011); Wold et al. (1987) is considered. PCA tries to reduce the dimensionality of the data by creating a smaller number of latent variables (components) as a linear combination of the original variables. Let us assume a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with already centered variables. PCA transforms the matrix $\mathbf{X}$ to a new coordinate system defined by

$$\mathbf{T} = \mathbf{X}\mathbf{P} + \mathbf{E}, \tag{1.5}$$

where $\mathbf{P}$ is an $d \times k$ loading matrix, $\mathbf{T}$ represents the $n \times k$ score matrix and $\mathbf{E}$ is an error matrix with the same dimensionality as $\mathbf{T}$. The jth column of the matrix $\mathbf{P}$ is computed in a way that the variance λ_j of the jth column of $\mathbf{T}$ is maximized. Furthermore, the different columns of $\mathbf{T}$ are supposed to be uncorrelated, which is equivalent to orthogonality constraints between the columns of $\mathbf{P}$. The number $k \leq min\,(n, d)$ is a pre-selected parameter, which determines the number or components of the new subspace.

A main disadvantage of feature extraction methods such as PCA or PLS is that new components are challenging to interpret. A set of k coefficients used to obtain new variables is returned for each original variable. Thus, if biomarker identification is the goal, feature extraction alone is not suitable. However, it can be – and in real-world application it often is – used as a classification (resp. regression) method in the first step, followed by some measure of relevance obtained from the fitted model.

For variable selection, one of the two types is usually applied – the threshold or a randomized approach. The threshold approach combines the loadings of the fitted model into one value for each variable based on a certain function. Then, the value is compared with a predetermined threshold to resolve the importance of the variable. The Variable Importance in Projection (VIP) (Wold et al., 1993; Favilla et al., 2013) or the Selectivity Ratio (SR) (Rajalahti et al., 2009a; Kvalheim, 2009) are frequently

used threshold methods. They are combining both loadings and explained variance of all components. For a closer description, see Chapter 2 and 4.

The randomized approach borrows the idea from permutation tests (Fisher, 1935; Rubin, 1980). It creates new artificial variables. They are constructed in a way that they do not carry any relevant information about the group structure. This might be done e.g. by bootstraping samples from the original variables and multiply them with a low constant as it is done by Uninformative Variable Elimination PLS (UVE-PLS), see Centner et al. (1996); Zerzucha and Walczak (2012). These artificial variables are then added to the original ones before fitting a statistical model (e.g. PCA or PLS). Thus, one can compare the loadings of the original variables with the distribution of the loadings of artificial ones and determine the threshold dividing biomarkers from noisy features. More details are described in Chapter 3.

Throughout the thesis, the focus is on feature extraction methods, since multivariate methods such as PCA or PLS are standard tools in metabolomics (Saccenti et al., 2014). However, e.g. embedded methods such as random forests were successfully applied in the field and its popularity is increasing (Chen et al., 2013; Truong et al., 2004).

1.4 Outlier identification methods

The history of anomaly detection, today referred to as outlier identification, goes back hundreds of years. It is well-known that outliers are present in almost all types of data, including metabolomics. There are two main reasons why outliers are and should be taken into account. *Firstly*, outliers often carry an important piece of information. They could refer to various reasons, from a different collection of the samples, different settings of a tool used for the analysis, or a structural defect up to frauds or medical problems. *Secondly*, even a small proportion of outliers in the data can distort the estimation of a statistical model (Huber, 2011; Abeel et al., 2009; Agostinelli et al., 2016). The simplest idea how to deal with outliers would be to remove them from the analysis. However, in order to do so, outliers must be identified correctly. This might be a challenging task, especially in higher dimensions (Filzmoser et al., 2008). Furthermore, if outliers are removed, the sample size of the dataset decreases which could affect the distribution of the data by e.g. underestimating the variance (Bellio and Ventura, 2005). A better strategy is to apply robust procedures, which generally downweight deviating observations instead of simply rejecting them.

PCA as an outlier detection tool

PCA, as described in the previous section, can also be used for outlier identification. There are, however, several issues connected with this approach. Classically, an eigenvector decomposition of the covariance matrix is employed in order to estimate the loadings in Equation (1.5). The sample variance which is being maximized and the sample covariance matrix are sensitive to outliers. Also Singular Value Decomposition (SVD) as a method to estimate PCA loadings is least-squares based, and thus sensitive to data outliers. So, in the presence of outliers, PCA becomes unreliable for identifying those outliers. As a solution, Croux and Haesbroeck (2000) investigated a robust PCA approach using the Minimum Covariance Determinant (MCD) (Rousseeuw, 1984) estimator for the covariance matrix. The involvement of the MCD indeed adds robustness properties to PCA. However, it cannot be used in situations when $d > n$, which is standard in metabolomics.

ROBPCA (ROBust method for Principal Components Analysis) (Hubert et al., 2005) combines the MCD estimator with the projection pursuit (PP) approach (Li and Chen, 1985; Hubert et al., 2002). The ROBPCA algorithm can be described in three steps. 1. PP is used for initial dimension reduction with at most $n - 1$ variables. This ensures applicability to high-dimensional data. 2. The initial covariance matrix is estimated and used for selecting the number of components k. In order to find a subspace with a good fit, the process is iteratively repeated. 3. The samples are projected on this subspace. Then, the robust location and the scatter matrix of the projected samples are computed. Finally, the eigenvectors of the scatter matrix which correspond to the biggest k eigenvalues give the robust principal components.

ROBPCA, apart from finding principal components with the biggest (robust) variance, has also another purpose – flagging outliers. To do so, it employs two distances measuring the outlyingness of the observations. The robust Score Distance (SD) and Orthogonal Distance (OD) for observation i are given by:

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{\lambda_j}}, \quad OD_i = ||\mathbf{x}_i - \mathbf{P}\mathbf{t}_i|| \,, \tag{1.6}$$

where $\mathbf{t}_i = (t_{i1}, \ldots, t_{ik})$ is the ith score vector, and λ_j the estimated variance of the jth component. SD represents the outlyingness as a distance of an observation in the PCA space relative to its center. It corresponds to the Mahalanobis distance of the observations in the score space. OD, on the other hand, measures the distance of each observation orthogonal to the PCA space. In order to classify an observation as non-outlying or

as outlying, two threshold values are employed. The threshold for SD is defined as the 0.975 quantile of the χ^2 distribution with k degrees of freedom, $\sqrt{\chi^2_{k,0.975}}$. As a cutoff for the orthogonal distances, it is suggested to take $\left(\hat{\mu} + \hat{\sigma}^2 z_{0.975}\right)^{3/2}$, where $z_{0.975}$ is the 0.975 quantile of the standard normal distribution. The values $\hat{\mu}$ and $\hat{\sigma}$ are estimated as median and median absolute deviation (MAD) of $OD_i^{3/2}$, respectively (Hubert et al., 2005). In the case that SD or OD exceed the cutoff values, the respective observation is considered as outlying. Furthermore, three types of outliers can be distinguished, outliers in the score space, in the orthogonal space, or in both spaces.

1.4.1 Cellwise outliers

Commonly, outlier identification has been carried out "rowwise", assuming that the observations are arranged in the rows of the data matrix. This means that if some method identifies an outlier, the complete observation is flagged. Robust statistical methods like ROBPCA would then typically downweight these observations, see Maronna et al. (2006). In contrast, "cellwise" outliers refer to a situation where single cells of the data matrix are deviating. 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 get downweighted. Cellwise outlier detection is a quite recent topic in robust statistics (Rousseeuw and Bossche, 2018), as well as is the development of robust estimators with cellwise outliers (Öllerer et al., 2016).

The identification of cellwise outliers is not an easy task. As an illustrative example, let us consider the two-dimensional case in Figure 1.6. Most observations follow a linear trend, but observations 1, 2 and 3 are clearly deviating. If the data would be examined univariately, one could conclude that observation 1 differs substantially in variable X1, whereas observation 2 differs in variable X2. However, observation 3 does not differ in either of the variables. Only when considering the bivariate information, both cells x_{31} and x_{32} would have to be selected as outlying. Let us assume there are more than two variables with relations to variables 1 and 2. It could turn out that cell x_{31} agrees with new variables but cell x_{32} does not. Thus, the latter cell would be classified as outlier. This shows that the relation between the variables should be considered in cellwise outlier identification.

In the following section two recent methods – DDC and cell-rPLR for the identification of cell-wise outliers will be described.

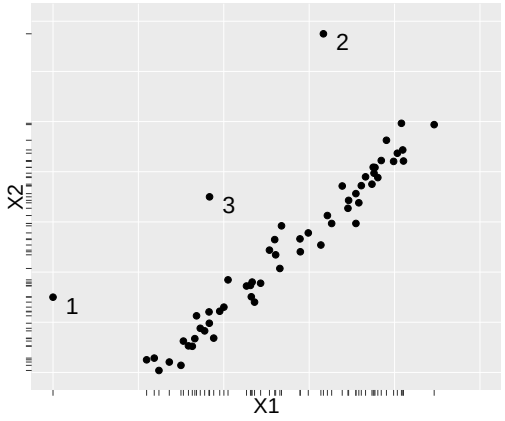


Figure 1.6: Illustrative example of bivariate outliers. Considering univariately X1 and X2, point 1 is deviating in X1, point 2 in X2. Point 3 is not deviating in of these variables, but is outlying in the bivatiate space.

Detecting deviating cells

Detecting deviating cells (DDC) is a recent algorithm by Rousseeuw and Bossche (2018). The method assumes that the majority of data cells are distributed according to a multivatiate normal distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with an unknown d-dimensional mean $\boldsymbol{\mu}$ and a positive semi-definite covariance matrix $\boldsymbol{\Sigma}$. However, some cells were due to various reasons altered or simply do not follow this distribution. In reality, it is recommended to transform all variables which do not follow Gaussian distribution to approximate Gaussianity (e.g. by Box-Cox power transformation (Box and Cox, 1964)). The DDC algorithm can be described in seven steps:

Step 1: *Standardization.* Each variable j of the data matrix $\mathbf{X}$ is centered and scaled by a robust estimation of location (*robLoc*) and scale (*robScale*).

$$z_{ij} = \frac{x_{ij} - m_j}{s_j}, \tag{1.7}$$

where

$$m_j = \text{robLoc}_i\left(x_{ij}\right) \text{ and } s_j = \text{robScale}_i\left(x_{ij} - m_j\right). \tag{1.8}$$

For details of the estimators, see Rousseeuw and Bossche (2018).

Step 2: *Univariate outlier detection.* The entries of the matrix $\mathbf{U}$ are defined as

$$u_{ij} = \begin{cases} z_{ij} & \text{if } |z_{ij}| \leq c \\ \text{NA} & \text{if } |z_{ij}| > c, \end{cases} \tag{1.9}$$

where c is a cutoff value selected as $\sqrt{\chi^2_{1;p}}$, with $p = 0.99$ as default.

Step 3: *Bivariate relations.* For each combination of variables $h \neq j$, a robust correlation *robCorr* is computed,

$$\text{cor}_{jh} = \text{robCorr}_i\left(u_{ij}, u_{ih},\right) \tag{1.10}$$

see Rousseeuw and Bossche (2018) for the detailed definition of the correlation measure. Only so-called connected variables h and j with relations

$$|cor_{jh}| > 0.5 \tag{1.11}$$

are kept. Furthermore, the slope of a robust regression line without intercept ($robSlope_i$) $b_{jh} = \text{robSlope}_i\left(u_{ij}|u_{ih}\right)$ of a regression of variable h on variable j is computed and will be used in the next steps.

Step 4: *Predicted values* are computed for each cell in the dataset. For variable j

$$\hat{z}_{ij} = G\left(\{b_{jh}u_{ih}; h \text{ in } H_j\}\right), \tag{1.12}$$

where for each variable j, H_j consists of variables satisfying the condition in Equation(1.11). G is a combination rule applied to these number which omit the NA values and is set to zero if H_j is empty. It is chosen as a weighed mean with weights $\omega_{jh} = |cor_{jh}|$.

Step 5: *Deshrinkage.* Since predictions $\hat{z}_{ij}$ tend to shrink the scale of the entities, $\hat{z}_{ij}$ is replaced by $a_j\hat{z}_{ij}$ for all i and j, where

$$a_j = robSlope_i\left(z_{i`j}|\hat{z}_{i`j}\right) \tag{1.13}$$

comes from regressing the observed $z_{i`j}$ on the predicted $\hat{z}_{i`j}$.

Step 6: *Flagging cellwise outliers.* Residuals are computed as

$$r_{ij} = \frac{z_{ij} - \hat{z}_{ij}}{robScale_{i`}\left(z_{i'j} - \hat{z}_{i'j}\right)}. \tag{1.14}$$

Then, cells which $|r_{ij}| > c$ are marked as anomalous, with c as in Step 2. Also, the matrix $\mathbf{Z}_{imp}$ is computed, which corresponds to the matrix $\mathbf{Z}$, but deviating cells and NA's are replaced by the predicted values $\hat{z}_{ij}$. The bigger the absolute difference between the elements of the matrix $\mathbf{Z}$ and $\mathbf{Z}_{imp}$ are, the more outlying are the corresponding cells.

Step 7: *Destandardize.* Lastly, the imputed matrix $\mathbf{Z}_{imp}$ is transformed back to an imputed matrix $\mathbf{X}_{imp}$. This is done by undoing the standardization in Equation (1.7).

The main output of the procedure is the matrix $\mathbf{X}_{imp}$ and the list of cellwise outliers. From this list, also rowwise outliers could be determined, see Rousseeuw and Bossche (2018).

cell-rPLR

The algorithm cell-rPLR which stands for cellwise outlier diagnostics using robust pairwise log-ratios is a novel approach which can be used for two goals: biomarker identification and cellwise outlier detection. The method can be summarized in four steps:

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 $j > k$. The index g refers to the group number, for $g = 1, \ldots, G$.

Step 2: Center and scale them robustly. Either according to the majority group or based on all observations. This gives values $\tilde{y}_{ijk}$, for all i, j, k.

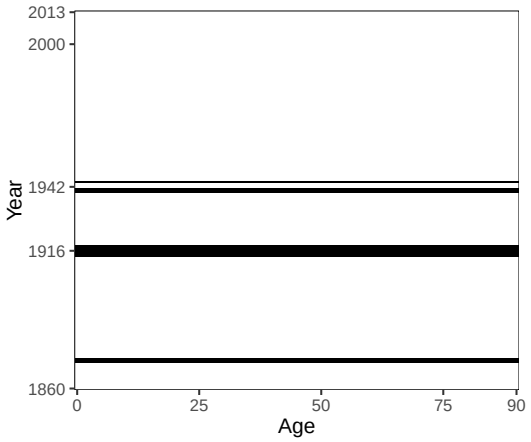
Step 3: Apply a weighting function to $\tilde{y}_{ijk}$, which yields weights w_{ijk}^{*}.

Step 4: Aggregate the weights to obtain the final weights w_{ij}, arranged in the weight matrix $\mathbf{W}$.

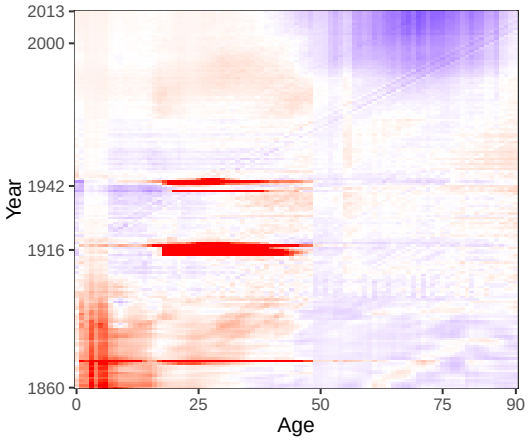
The final weights are in the range $[-1, 1]$ and they express the degree of outlyingness of single cells. Weights around zero represent non-outlying values, and weights closer to -1 or $+1$ represent potential outliers. The cell-rPLR algorithm is more introduced in more detail in Chapter 4.

Mortality dataset

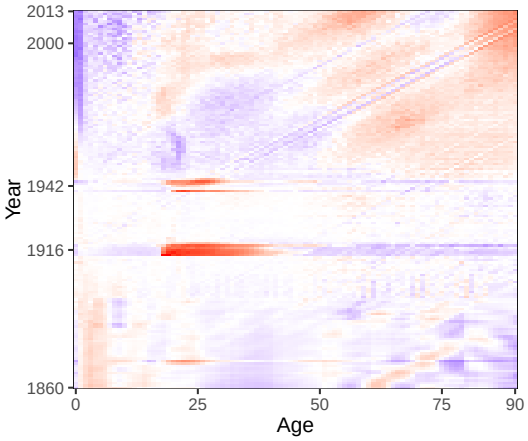
As a non–metabolomic example, the mortality dataset is analyzed by ROBPCA, DDC and cell-rPLR. The data can be obtained from the R package *cellwise* (Raymaekers et al., 2018). It describes the mortality rates of males in France between the years 1860 and 2013, for ages between 0 and 91. Each row of the dataset corresponds to a certain year and each column to an age. Figure 1.7(a) shows the outlier identification results of ROBPCA. It identifies the whole observations as outlying. The years of the first and the second world wars as well as the years of the Prussian war were identified. Figure 1.7(b) and 1.7(c) is a heatmap of standardized residuals of DDC and of the result of the cell-rPLR algorithm with the Tukey biweight function, respectively. The red colour



(a) ROBPCA



(b) DDC



(c) cell-rPLR

Figure 1.7: Mortality dataset. Heatmap of outliers for three methods. (a) ROBPCA, (b) standard residua of DDC and (c) cell-rPLR with Tukey biweight function

corresponds to a "positive outlier". Positive outliers represent a cell value higher than expected. On the other hand, blue color is a visualization of a "negative outlier" – a value lower than expected. Both 1.7(b) and (c) provide more insight into the data than ROBPCA, since cellwise outliers are flagged. Both methods correctly identified all three wars, but they are revealing more details. One can see that an increased mortality rate relates only to the group of people between approximately 18 and 45 years, presumably soldiers. DDC reveals that child mortality was very high in the past, which is not so clear from cell-rPLR, where only slightly increased values for newborns aged zero to one are visible. On the other hand, the cell-rPLR method adds that France had seen a dramatic decrease of the newborn mortality rate in the last 50 years. The possible

reason of the difference might be that DDC does not take into notice the compositional structure of the data. Furthermore, a decrease in newborn mortality is unprecedented, and thus DDC might have not found any correlated variables. Thus, the prediction for this period would not be precise.

1.5 Outline

This thesis introduces two novel methods for feature selection and identification of cellwise outliers. These methods aim to be used mainly, but not necessarily only in the field of metabolomics. A crucial part of the thesis is the application of the algorithm to real-world datasets. All developed procedures and graphics were created with the software environment R (R Core Team, 2018).

Chapter 2 describes some problems and issues connected with the data coming from omics disciplines which are measured either by Mass Spectrometers or Nuclear Magnetic Resonance Spectrometers. A proper pre-treatment of the data is crucial in order to understand the biological information accurately. The impact of several commonly used normalization, scaling and transformation methods is measured based on two most common objectives in omics – classification and feature selection. In addition, log-ratio transformations are considered and compared to the other approaches. Recommendations for appropriate pre-treatment methods are given. Furthermore, possible explanations for methods with poorer performance are provided.

J. Walach, P. Filzmoser, and K. Hron. Data normalization and scaling: Consequences for the analysis in omics sciences. In: J. Jaumot, C. Bedia, and R. Tauler (eds.) Comprehensive Analytical Chemistry. Data Analysis for Omics Sciences: Methods and Applications. Elsevier, Amsterdam, The Netherlands, pp. 165-196, 2018.

Chapter 3 presents the biomarker identification method rPLR. Thanks to robust estimation of the variance, rPLR is highly robust against data outliers. Additionally, the method can be applied in cases of unequal group sizes. A simulation study as well as a real world dataset were analyzed in order to test the performance of rPLR. Focus is given on the case when deviating observations or cells are present in equal and unequal data structure.

J. Walach, P. Filzmoser, K. Hron, B. Walczak, L. Najdekr. Robust biomarker identification in a two-class problem based on pairwise log-ratios. Chemometrics and Intelligent Laboratory Systems, 171, pp. 277-285, 2017.

Chapter 4 introduces the algorithm cell-rPLR for cellwise outlier identification. Data outliers can carry very valuable information and often are the most informative for the interpretation. Pairwise log-ratios between the variable values form the elemental information for the algorithm, and the aggregation of appropriate weights results in outlyingness information. Cell-rPLR can also be used for biomarker identification, particularly in presence of cellwise outliers. Real data examples and simulation studies underline the good performance of this algorithm in comparison to alternative methods.

J. Walach, P. Filzmoser, Š. Kouřil. Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log-ratios. Submitted for publication, 2018.

Chapter 5 presents functionality of two R packages, based on the rPLR and cell-rPLR methods described in Chapters 3 and 4.

Unpublished.

CHAPTER 2

Data normalization and scaling: consequences for the analysis in omics sciences

Abstract: Nowadays, the use of different types of measurement devices such as Mass Spectrometers or Nuclear Magnetic Resonance Spectrometers are standard in "omics" disciplines. Such a device produces high-dimensional data, but they cannot immediately undergo a statistical analysis because the measured samples and features are usually not directly comparable. This is due to different sample volume, different feature abundance, or different error variance implying heteroscedasticity. Thus, a proper pre-treatment of the data is crucial in order to understand the biological information accurately. The impact of several commonly used normalization, scaling and transformation methods is reviewed, and the methods are tested based on the two most important objectives in this context – classification and feature selection analysis. Recommendations for appropriate pre-treatment methods are provided, and possible explanations for methods with poorer performance are given.

Keywords: Pre-treatment methods, Normalization, Scaling, Transformation, Classification, Feature selection, Omics, Metabolomics

2.1 Introduction

Quantitative analyses in the "omics" sciences are important in order to understand the chemical and biological relationships. Nowadays, Nuclear Magnetic Resonance (NMR) Spectrometers and different types of Mass Spectrometers (MS) are the main tools to analyze biological samples due to their superior detection sensitivity (Rainville et al., 2014). NMR and MS produces a big amount of data in form of abundances of certain components of the biological sample. To obtain, understand and interpret the important information from the produced data, it is desirable to apply bioinformatical or statistical techniques to the data. However, there are several steps before the final data can be used as an input of a statistical analysis method. Firstly, the biological experiment needs to be designed. Biological samples of, e.g., blood, plasma, cells or urine are collected, prepared and analyzed for instance by MS. Afterwards, certain pre-processing (Van Der Werf et al., 2005; Shurubor et al., 2005) steps are performed to reflect the concentrations or intensities of ions or m/z values. Such data could be used as an input for the statistical analysis. However, a further pre-treatment step is crucial to achieve relevant and unbiased conclusions. The pre-treatment methods make sure that the data are converted in a way that all the samples and all the variables can directly be compared. Sometimes, the samples cannot be directly compared due to different volumes. For example, if twice of the amount of the same sample would be analyzed, one could expect around twice as high abundances for all variables. On the other hand, the measured features might have differences in order of magnitude between the measured concentrations. This, however, does not generally mean that the variables with higher average abundances are more important than the another ones. Another point important for a pre-treatment method is to reduce noise and to focus on the important information contained in the data. Considering two groups in the data (e.g. controls and patients), certain variables might have increased abundances for one group. This is called *induced biological variation*. Often, the goal of the analysis is to find induced variation in order to understand biological processes in the samples. On the other hand, there are several types of "unwanted variation". There are three types of such a variation. Firstly, biological variation describes within-subject variability, e.g. two samples of the same tissue are analyzed, and this will lead to abundances that will slightly differ. Technical variation is caused by errors in a Mass Spectrometer or by analytical errors. Many statistical methods assume that the errors fluctuate around zero with constant variation. Unfortunately, often with increasing concentration, the variation of noise also increases. Such a situation is called heteroscedasticity. The goal of some of the pre-treatment methods is to convert this to

homogeneity, i.e. to make the error variances comparable.

The pre-treatment methods can be separated into three categories: normalization and transformation methods, and scaling methods. From a general point of view, normalization and transformation methods as described in Section 2.2 mainly deal with the differences between the samples, and the transformation methods suppress heteroscedascity. Scaling methods, described in Section 2.3, try to make the features comparable. It is also possible to combine different types of methods in order to achieve even better results.

The main task in the analysis of omics data is to understand biological information in the data. From a statistical point of view, classification analysis is one of the goals. If the data are consisting of groups of e.g. controls and patients, the accurate prediction of new samples is desirable. To understand the processes in the human body or in other organisms, the interpretation of the model is necessary. In the two-group setting, the information about important features is one of the main tasks in omics disciplines. In metabolomics, the problem is called biomarker identification, while in genetics this is called fold changes problem, where it is examined for a feature, how many times the average concentration in one group is higher/lower than for the other group. In statistics, this is often referred to as the feature selection problem. Section 2.4 analyzes the impact of pre-treatment methods on publicly available real world data sets in terms of classification and feature selection analysis. As an example of omics disciplines, the data sets are originating from the metabolomics field. The final Section 2.5 discusses and summarizes the main findings, and provides some overall recommendations.

2.2 Normalizations and transformations

The goal of normalization methods is to make the data values of the measured samples comparable among each other for the subsequent statistical analysis. Samples often cannot be compared directly because of differences in their volumes – the so-called size effect. Typical examples are metabolomic data derived from MS or NMR. If the analyzed materials are originating, e.g., from urine samples, the concentrations–which are directly related to the volumes–might differ by a factor of 10-15 (Warrack et al., 2009; Tsuchiya et al., 2003). The reasons are manifold: different fluid consumption of the patient, different drug or toxin intake, different treatment methods, or different physiological factors. All these conditions might lead to an increase or decrease of the concentrations of urine. If the difference in concentration levels is not taken into account when analyzing the data (Webb-Robertson et al., 2005), misleading or even incorrect conclusions might be the consequence. These mentioned shortcoming are not only problematic for metabolomics

data, but also other omics disciplines suffer from similar issues.

Next to normalization techniques, data transformation is another possibility of data pre-treatment. There is no clear distinction between normalization and transformation methods, however, the names correspond with the use in literature. In this section, we describe two types of transformations: non-linear and log-ratio transformations.

In the following we will assume a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with n samples in the rows and d variables in the columns.

2.2.1 Normalization methods

Normalization methods could be separated into two groups: (a) normalization methods based on specifically measured features (e.g. metabolites), and (b) data-driven methods. The group (a) consists only of normalization to internal standard. It uses an expert information about a variable or set of variables which can be used as baseline for normalization of other variables. Further, data-driven methods normalize the data based on statistical methods.

- **Normalization to internal standard.** This method is sometimes called normalization to a "housekeeping" variable. The procedure works as follows: A reference variable is chosen, and each other variable is divided by its intensity. This kind of normalization is one of the most frequently used normalization methods in the field of analytical chemistry. If "housekeeping" variables are present in the data, internal standard normalization should always be recommended when experimental data are acquired. A specific example is an analysis of urine samples in metabolomics. Since each urine sample might have different concentration of water, the intensities of the samples are also different. Then, for example, the metabolite Creatinine can be used as a reference, since it is assumed that Creatinine is constant in urine, and thus its concentration is directly equivalent to the urine concentration (Bolstad et al., 2003). This assumption holds under normal conditions, however, this assumption does not always hold (Garde et al., 2004), and in that case one could expect biased results.

  One should keep in mind that the use of internal standard normalization is not limited only to the case of metabolomic analyses of urine sample, but it should be rather considered for each experimental data set from analytical chemistry (Sysi-Aho et al., 2007; Skoog et al., 2017; Katajamaa and Orešič, 2007).

- **Total sum normalization** (TSN): This method is sometimes also called *normalization to a constant sum* (Craig et al., 2006), *constant sum normalization* (Giraudeau

et al., 2014) or *total spectral area normalization* (Saccenti, 2017), and it is another commonly used normalization method in the omics sciences (Filzmoser and Walczak, 2014). The goal of TSN is to make the samples comparable by forcing the sum of each sample to be equal to a constant, most commonly to 1 or to 100 (%). After TSN, the variables represent fractions or percentages of the whole sample. This is achieved by dividing each sample by the sum of the values of the sample. Formally, for the ith sample $x_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$, for $i = 1, \ldots, n$, TSN is defined as:

$$x_i^{TSN} = \left[\frac{x_{i1}}{t_i}, \frac{x_{i2}}{t_i}, \ldots, \frac{x_{id}}{t_i}\right], \qquad \text{where } t_i = \sum_{j=1}^{d} x_{ij}. \tag{2.1}$$

Equation (2.1) normalizes each sample to a sum of 1. By multiplying all elements of x_i^{TSN} by a constant c (e.g. $c = 100$), a total sum of c would be achieved. The situation when each observation adds to a constant sum is sometimes referred as "closure" or "closed" data sets (Filzmoser and Walczak, 2014). Despite the frequent use of TSN, this normalization can lead to incorrect conclusions (Filzmoser and Walczak, 2014). In order to allow one or several measured intensities to increase in a constant sum scenario, others will have to decrease, even though this may not necessarily correspond to the given problem. Spurious correlations may result, indicating forced relations among the variables. A further disadvantage of TSN normalization is the fact that the differences are redistributed among all variables equally, thus it creates artificially increased differences in other variables. Lastly, the lack of robustness is also an issue, since only one outlying cell in the data matrix can disturb the normalization of the whole sample.

- **Probabilistic quotient normalization** (PQN) (Dieterle et al., 2006) is a popular linear normalization method, which is mainly used to deal with the so-called size-effect. It is assumed that the size-effect influences the whole sample and that biological changes between samples in measured concentrations affect only some parts of the spectra. As a practical consequence of this, PQN assumes that at most half of the variables show biological changes. This is a limitation of PQN, since if there are biological changes in more than half of the measured concentrations, PQN cannot be correctly applied.

  The algorithm for PQN is as follows:

  1. Calculate/determine a reference spectrum. There are two possibilities in this step. Firstly, one can determine "the golden spectra" as a reference. However, in real world situations this is hardly ever known in advance. Secondly, the reference

spectrum might be computed from the data set itself. Typically, the median or mean spectrum of the control samples is used, where the median is recommended due to its robustness properties,

$$x_j^{ref} = \text{median}\,\{x_{1j}, \ldots, x_{nj}\} \quad \text{for} \quad j = 1, \ldots, d.$$

2. Compute the ratios of all variables with the corresponding reference spectrum, and take the median of these values,

$$x_i^* = \text{median}\left\{\frac{x_{i1}}{x_1^{ref}}, \ldots, \frac{x_{id}}{x_d^{ref}}\right\} \quad \text{for} \quad i = 1, \ldots, n.$$

This estimates the abundance of the size-effect for each observation.
3. Divide all samples by the corresponding median. The division normalizes the samples to the same concentration level,

$$\mathbf{x}_i^{PQN} = \left\{\frac{x_{i1}}{x_i^*}, \ldots, \frac{x_{id}}{x_i^*}\right\} \quad \text{for} \quad i = 1, \ldots, n.$$

- **Cyclic Loess normalization** (cLOESS), where Loess stands for Locally weighted scatterplot smoothing (Cleveland and Devlin, 1988). The algorithm for Cyclic Loess is as follows:
1. Choose two samples x_i, x_j from the dataset, and log-transform (base 2) the values component-wise, resulting in $y_i = \log_2(x_i)$ and $y_j = \log_2(x_j)$.
2. Compute their difference $m^{ij} = y_i - y_j$ and average $a^{ij} = \frac{(y_i + y_j)}{2}$.
3. Use m^{ij} as response and a^{ij} as explanatory variable, and fit with Loess. The fitted values are $\hat{m}_{ij} = (\hat{m}_1^{ij}, \ldots, \hat{m}_d^{ij})$.
4. Update y_i and y_j with the fitted values components-wise according to $y_i \leftarrow y_i + \hat{m}^{ij}/2$ and $y_j \leftarrow y_j - \hat{m}^{ij}/2$.
5. Repeat steps 1.-4. for each possible combination of samples.
6. Repeat steps 1.-5. until convergence (until the values stabilize).

Thus, this normalization builds on log-transformed Bland-Altman plots, called MA-plots (Altman and Bland, 1983), which plot the components of m^{ij} against those of a^{ij}. Then, Locally weighed polynomial regression, also called Locally weighted scatterplot smoothing (Loess) Cleveland and Devlin (1988) is used, which is able to model non-linear relationships, see Cleveland and Devlin (1988) or Dudoit et al. (2002), and the log-transformed values are updated. This is done by cycling through all sample pairs, and repeating the procedure until convergence, where

usually two iterations are sufficient (Kohl et al., 2012). The idea behind Cyclic Loess is that if two samples are comparable and well normalized, the difference between all features of the samples will be around 0. As a first step, the logarithm is applied on the data which helps to deal with the heteroscedasticity. However, because of the logarithm, zero measurements need to be replaced, see Section 2.4.2.

- **Contrast normalization** (Contrast). This method has been described in Åstrand (2003). Similar to Cyclic Loess, this method is based on MA-plots (Altman and Bland, 1983). Firstly, the data matrix is log-transformed to deal with the heteroscedasticity. Then, the data matrix is multiplied by a specific orthonormal transformation matrix $\mathbf{T}$ (Åstrand, 2003), thus it is moved to the so-called contrast space. The use of the contrast space extends the idea of MA-plots to the multidimensional case, because the first column of the contrast transformed intensities is used as a predictor for the rest of the variables. A set of Loess regressions is fitted similarly as in Cyclic Loess, but in order to achieve robustness, a re-descending M estimator (Hampel et al., 1981) with the bisquare weight function is used. The Euclidean distances of the estimated values from Loess regression and the contrast values are taken as a set of robust weights in each iteration. The robust weights are independent from the orthonormal transformation. Lastly, the data are back-transformed to the original input space. The advantage compared to Cyclic Loess is that Contrast normalization is not that computationally intensive.

- **Quantile normalization** (Quantile) (Bolstad et al., 2003). The objective is to force the samples to have the same distribution of the feature intensities. Quantile normalization can be described in two steps. Firstly, the values are sorted for each variable separately in ascending order. Secondly, the means of the same quantiles across all features are computed, i.e., the mean of the lowest values, the mean of the second lowest values, etc. These mean values are assigned to all variables for the same corresponding quantiles, i.e., the lowest value of each variable is replaced by the mean of all lowest values of all variables, etc. After the normalization, all variables consist of the same values, but the order of these values will in general be different for different variables. Thus, when comparing the distribution of any two variables in a QQ (quantile-quantile) plot, all points are arranged on a straight line, indicating identical distributions.

- **Baseline normalization**. The basic idea is to choose a baseline spectrum and adjust the measured samples based on the baseline. Common approaches to choose a proper baseline are to take the mean or the median spectrum of all samples (or

of the samples of a group – if sub-groups exist and are known), to take the "golden spectrum", or a random sample from the data. Then one tries to map the samples to the baseline. Three mapping methods are common:

- **Linear baseline normalization** (lBase) (Bolstad et al., 2003) computes a scaling factor to map each sample to the baseline sample. The scaling factor is computed separately for each sample as a ratio between the mean intensity of the baseline and the mean intensity of the spectrum. The values of each sample are then multiplied by the corresponding scaling factor.
- **Non-linear baseline normalization**(nlBase) (Li and Wong, 2001). This method is preferable if non-linear relations between the baseline and the samples are assumed. The ordered values of a sample are plotted against the ordered values of the baseline, and a curve (smoothing spline) is fit. This curve defines the mapping between the sample and the baseline, and is used to correct the sample values. This can also be combined with feature selection by comparing the rank of the intensities of the baseline and the rank of the intensities of the samples. Only similarly ranked intensities are kept for curve fitting. The feature selection process is iterative and is done separately for each sample, and thus different features may be selected for different samples.
- **Cubic splines normalization** (cSplines) (Workman et al., 2002) also assumes a non-linear relation between the baseline and the samples. The idea is to force all features to have similar distributions. Thus, the method is in this aspect similar to Quantile normalization. The normalization is performed by fitting cubic splines between the baseline as a response and each sample separately as explanatory variable, in a way that a set of evenly distributed quantiles between the two are fitted by smooth cubic splines. The fit is done several times while choosing slightly different subsets each time. Then, all fitted splines are used to fit the parameters of natural cubic splines. Values which are between two quantiles are adjusted by interpolation based on the neighboring quantiles.

2.2.2 Non-linear transformations

The main objective of non-linear transformations is generally to deal with heteroscedasticity, to make skewed distributions more symmetric and thus to approach normality of the data. From a biological point of view, the relations among the variables can also be multiplicative and not just additive as classically expected. Then, a transformation

is a necessity to examine the data properly. Most often used transformations are the logarithmic transformation and the power transformation.

Both transformations mentioned in this section reduce high intensities much more than small intensities, which can be even amplified. Thus, the data range after a transformation is reduced. This is a similar principle as with scaling methods. However, since no real scaling factor is used, transformations are sometimes called "pseudo-scaling" (van den Berg et al., 2006). The scaling effect of transformations is often not strong enough to entirely deal with differences in orders of magnitude, so it might be desirable to use some scaling method after transformation.

- **Log-transformation** (Log). The logarithm of the data values is computed to remove heteroscedasticity, see Kvalheim et al. (1994). The log transformation cannot deal with zero-values, since $\log(0) = -\infty$. This also means that values close to zero will be emphasized (with negative sign), and thus low measured intensities will be spread on a big range of values. The zero problem can be avoided by imputing the zero values as described in Section 2.4.2.

- **Power transformation** (Sqrt) (Bickel and Doksum, 1981). The data values are powered by some chosen constant, such as 1/2 for the square root. This transformation does not suffer from zero problems, and it also has the goal of removing heteroscedasticity. Thus, the power transformation is often used if zeros are present, because it gives a similar transformation pattern as the log-transformation. The disadvantage over log-transformations is that it cannot change multiplicative to additive effects.

2.2.3 Log-ratio transformations

Data consisting only of strictly non-negative values which are part of a whole are called compositions. The term *compositions* is used in compositional data analysis (CoDa), seePawlowsky-Glahn and Buccianti (2011). Compositions are often expressed in the form of percentages, probabilities, frequencies or – as in omics – concentrations or counts per unit. Compositions are often "close-to-closure". This means that the sum over all compositions (features) is almost equal for all samples. This is, however, not a requirement for the definition of compositional data. Data which are far-from-closure can also be viewed as compositional. The main difference from a "traditional" and compositional point of view on the data is the fact that the latter assumes that the important information is carried between the ratios of the variables rather than in the absolute values. To extract the information between the ratios, so-called log-ratio

transformations were introduced (Aitchison, 1986). Log-ratio transformations make use of the log-ratios $\log(x_{ij}/x_{il})$ for any variable pair with index j and l (here for the ith observation). The resulting log-ratio approach also argues that the sample space of compositions is the so-called simplex space, which is a subspace of the classical Euclidean space (Pawlowsky-Glahn et al., 2015). The goal of log-ratio transformations is to move the data from this simplex to the usual Euclidean space, such that standard statistical methods can be used.

There are three important principles which should be fulfilled for a compositional data analysis: scale invariance, subcompositional coherence, and permutation invariance. Scale invariance follows the rule that only ratios between the compositions are important. Transformations which are based on log-ratios will be invariant to scaling with a factor s, since $\log\left(\frac{s\cdot x_{ij}}{s\cdot x_{il}}\right) = \log\left(\frac{x_{ij}}{x_{il}}\right)$. Note that normalization methods such as TSN or PQN, applied beforehand, will have no effect on the results if the scale invariance principle is fulfilled. Subcompositional coherence states that the results should not be in contradiction if the whole composition is examined or if any subcompositions is used. Lastly, permutation invariance means that the compositional data analysis must not dependent on the order of the compositional parts.

In the following we list two log-ratio transformations which are isometric, meaning that they preserve distances. There are also other well-known log-ratio transformations, such as the additive log-ratio transformation, which do not have this important property (Pawlowsky-Glahn et al., 2015).

- **Centered log-ratio** (clr) transformation (Aitchison, 1986; Pawlowsky-Glahn et al., 2015): The ith observation $x_i = [x_{i1}, \ldots, x_{id}]$ is transformed to

$$x_i^{clr} = [x_{i1}^{clr}, \ldots, x_{id}^{clr}] = \left[\log\left(\frac{x_{i1}}{g(x_i)}\right), \ldots, \log\left(\frac{x_{id}}{g(x_i)}\right)\right], \tag{2.2}$$

  where $g(x_i) = \sqrt[d]{\prod_{j=1}^{d} x_{ij}}$ is the geometric mean of the ith observation, for $i = 1, \ldots, n$. Thus, clr transformed data have the same dimension d as the original dataset, however, the components sum up to zero, $x_{i1}^{clr} + \ldots + x_{id}^{clr} = 0$. This means that clr transformed data do not have full rank d, which could create problems for methods like discriminant analysis, where a covariance matrix with full rank is required. Also for some robust statistical methods this is a prerequisite. On the other hand, the components of clr transformed data have a straightforward interpretation in terms of a dominance of the corresponding compositional part on an average behavior (geometric mean) of the values in the composition. Note that clr transformed data cover log-ratio information of all different pairs of variables:

for example, the first component can be written as

$$x_{i1}^{clr} = \frac{1}{d}\left(\log\left(\frac{x_{i1}}{x_{i2}}\right) + \ldots + \log\left(\frac{x_{i1}}{x_{id}}\right)\right).$$

- **Isometric log-ratio** (ilr) transformation (Egozcue et al., 2003): The clr transformation mapped the compositions from the simplex to a $(d-1)$-dimensional hyperplane in the d-dimensional real space. Now, ilr is a class of transformations which builds an orthonormal basis in this hyperplane and expresses the compositions in this orthonormal basis. Thus, the ith compositional observation $x_i = [x_{i1}, \ldots, x_{id}]$ is transformed to $x_i^{ilr} = [x_{i1}^{ilr}, \ldots, x_{i,d-1}^{ilr}]$, with only $d-1$ components, and since they are expressed in an orthonormal basis, they will be called *ilr coordinates*. There are infinitely many possibilities to set up such an orthonormal coordinate system, and one specific choice are *pivot coordinates*, where the jth component of x_i^{ilr} is defined as

$$x_{ij}^{ilr} = \sqrt{\frac{d-j}{d-j+1}} \cdot \log\left(\frac{x_{ij}}{\sqrt[d-j]{\prod_{k=j+1}^{d} x_k}}\right), \text{ for } j = 1, \ldots, d-1, \tag{2.3}$$

see Fišerová and Hron (2011). By construction, only the first component x_{i1}^{ilr} includes information of the first variable x_{i1}, but such information is not contained in any other component. Moreover, one can show that $x_{i1}^{ilr} = \sqrt{\frac{d}{d-1}} x_{i1}^{clr}$, i.e. the first ilr component is proportional to the first clr component, and thus it contains all relative information (in terms of log-ratios) of the first variable to the remaining variables in the composition. This makes the interpretation of this first ilr component very unique, namely as the dominance of the first variable to an average behavior of the other variables in the composition. Note that for clr it was not possible to extract all relative information into one clr component, since the geometric mean is involved in all the variables. If the interpretation is required for another variable, then this variable needs to be reordered to the first position in the data set, pivot coordinates need to be computed, and again the first coordinate represents all relative information about this variable of interest.

The ilr transformation will not be used in the subsequent analyses, because the methods employed can cope with the zero constraint of the clr transformed data. For other methods, however, this can be a valuable alternative.

2.3 Scaling

The aim of scaling as a pre-treatment method is to deal with different scale among different variables. Thus, its goal is to adjust the variance of each variable and to make all variables similarly important, and to deal for example with heteroscedasticity. Scaling methods divide each variable by a so-called scaling factor, which is in general not the same for each variable. A frequent problem with scaling is the possible inflation of small values, which could imply that also measurement errors are increased. There are two types of scaling methods: the first type uses size measures, such as mean or median, while the second type uses the data dispersion (e.g. standard deviation or median absolute deviation).

Scaling is almost always applied after **centering** the data. Centering, as well as scaling, is done for each variable separately. Before centering the data, the concentrations of each metabolite scatter around the center of the distribution (i.e. mean, median). Centering levels the central values to the same value of zero. Thus, the abundance (meaning low or high values) of the original variables is not important anymore and should not affect further analyses of the data.

In the following, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ denotes the arithmetic mean of the jth variable, and $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$ stands for the empirical standard deviation of the jth variable, for $j = 1, \ldots, d$.

Figure 2.1 illustrates the effect of centering and scaling. Here, the arithmetic means of the variables were used for centering, and the empirical standard deviations for scaling. For this reason, the medians shown in the boxplots and the interquartile ranges represented by the lengths of the boxes are not perfectly matching for the different variables.

- **Autoscaling** (Auto) or *Unite variance scaling* (Jackson, 2005) is the simplest scaling method used in omics disciplines. The aim of the method is to normalize the variables in a way that each of them has a mean of zero and a variance (thus also standard deviation) equal to one. This is achieved by subtracting from the values of each variable the mean and dividing by the standard deviation of that variable,

$$x_{ij}^{Auto} = \frac{x_{ij} - \bar{x}_j}{s_j}. \tag{2.4}$$

  After autoscaling, the analysis of the data by many multivariate statistical methods (e.g. LDA, PCA, PLS, ...) will not be based on covariances but on correlations.

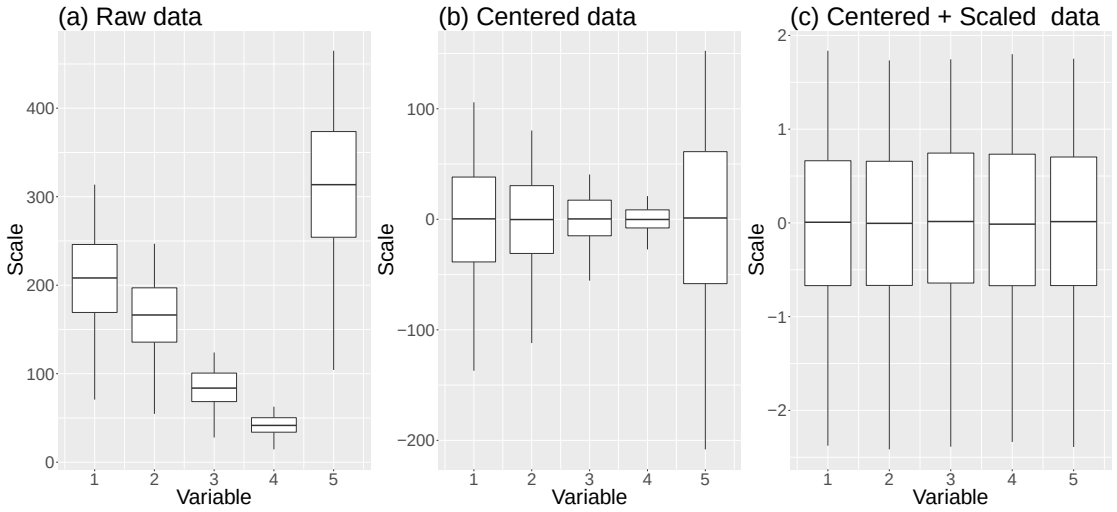


Figure 2.1: Effects of centering and scaling. Boxplots for five variables are shown. Plot (a) shows raw unprocessed data. The data for each variable scatter around different central values. The variability is also different. Plot (b) show mean-centered data, and now the values from all variables scatter around zero. Plot (c) shows centered and scaled data, resulting in comparable scales of the variables.

Autoscaling takes all features as equally important, since all the features will have comparable scale. The method as defined here is not robust, and thus it can be influenced by outliers. However, this could be solved by replacing the arithmetic mean by a robust counterpart, e.g. the median, and the empirical standard deviation by a robust scale estimator, such as the median absolute deviation. Another problem with autoscaling is the fact that measurement errors can be amplified. This is the case if the estimated standard deviation is small; then it increases all the abundances for a variable through the division in Eq. (2.4).

- **Pareto scaling** (Pareto) (Eriksson, 1999; Kubinyi, 1994) is a modification of autoscaling. After centering the variables, the square root of the standard deviation is used as a scaling factor,

$$x_{ij}^{Pareto} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j}}. \tag{2.5}$$

The advantage of this approach is that the scaled data remain closer to the original data because the scaling effect is less intense. The measurement noise is not amplified as much as in the case of autoscaling. Also large fold changes will not be as important as before, but very large fold changes will still play a major role in normalization. Since non-robust estimators for centering and scaling are involved,

there might be an effect of outliers.

- **Range scaling** (Range) (Smilde et al., 2005) starts with centering the variables first. Denote $x_{j_{max}}$ and $x_{j_{min}}$ as the maximum and the minimum of all values of the jth variable, respectively, for $j = 1, \ldots, d$. The range is defined as the difference $x_{j_{max}} - x_{j_{min}}$, and it is a measure of scale of the jth variable. Since this scale measure is based on the most extreme data values, it is very sensitive to outliers. The range scaled data are defined by

$$x_{ij}^{Range} = \frac{x_{ij} - \bar{x}_j}{x_{j_{max}} - x_{j_{min}}}, \tag{2.6}$$

  and the aim is to make each variable equally important for the subsequent analysis. The natural minimum for measuring intensities, e.g. from mass spectra, is zero, and thus only the maximum value is often of importance. Similar as in autoscaling, error inflation is also a problem for range scaling.

- **Level scaling** (Level) (van den Berg et al., 2006) belongs to the first type of scaling methods since it uses the estimation of location instead of scale (spread) as a scaling factor,

$$x_{ij}^{Level} = \frac{x_{ij} - \bar{x}_j}{\bar{x}_j}. \tag{2.7}$$

  The method transforms the changes in intensities to changes relative to average intensities by using the arithmetic mean as a scaling factor. After the application of level scaling, the values are interpreted as proportional changes compared to the mean intensity. Level scaling is sensitive to outliers, which can be avoided by using the median instead of the arithmetic mean. The method is usually used if the focus is on biomarker identification. Again, a possible inflation of errors is a disadvantage of the method.

- **Vast scaling** (Vast) (Keun et al., 2003) is a shortcut for variable scaling. It is a modification of autoscaling which focuses on the stability of the variables. This method makes use of the coefficient of variation (cv), which is computed for the jth variable as $cv_j = \frac{s_j}{\bar{x}_j}$. Vast scaling is then defined as

$$x_{ij}^{Vast} = \frac{x_{ij} - \bar{x}_j}{s_j} \cdot \frac{1}{cv_j}. \tag{2.8}$$

  Thus, the data are autoscaled, and then the cv is used as an additional scaling factor. The cv stabilizes the variables in a way that it highlights higher interest for features with small relative standard deviation. On the other hand, it decreases

the importance of features with large relative standard deviation. The method can be used in unsupervised but also in supervised settings. Thus, group information of the samples can be incorporated. In a supervised setting, the *cv* is computed separately for each group, and its mean is taken as a scaling factor.

- **Variance stabilization normalization** (VSN) (Huber et al., 2002) is a method combining normalization methods with a stabilization of the feature variances. The literature offers several possible versions of the method. Here, the method of Kohl et al. (2012) will be described. The goal of VSN is to make the variance constant over the entire data range. Firstly, the between-sample variation is reduced by linearly mapping all samples to the reference sample (the first sample in the dataset). Then, the adjustment of the variance of the data is performed. Similar as in Vast scaling, the coefficient of variation is examined. The method assumes a relationship between the standard deviation and the mean, i.e., with increasing mean one could expect an increase of the variance. VSN assumes a quadratic relation between the ratio of mean and standard deviation. However, since the lower limit of the measurement is zero, the variance within small values will not decrease anymore but stays more or less constant. Thus, the coefficient of variation increases. To deal with this, VSN uses the inverse hyperbolic sine function. The function asymptotically follows the logarithmic function for large values, which removes heteroscedasticity. However, for small intensities the linear relationship is kept, and thus the variance is unchanged.

2.4 Practical aspects of the methods

The main focus in this section is on the ability of the different normalization and scaling methods to correctly classify data into given subgroups (e.g. healthy people versus diseased), and to accurately select features. Basically, these are different goals, and a method with a high accuracy in feature identification must not necessarily be precise for classification, and vice versa. In the following, the different methods and combinations thereof will be tested on some data sets.

Many of the previously mentioned methods are implemented in software packages. Table 2.1 summarizes R (R Core Team, 2018) functions available in existing packages or in a newly developed package published on Github.

Method	R package::function
TSN	KODAMA::normalization, method = 'sum'
PQN	KODAMA::normalization, method = 'pqn'
cLOESS	+limma::normalizeCyclicLoess
Contrast	+affy::normalize, method = 'contrast'
Quantile	preprocessCore::normalize.quantiles
lBase	*pretreatment::LinearBaseline
nlBase	*pretreatment::NonLinearBaseline
cSplines	+affy::normalize.qspline
Log	base::log
Sqrt	base::sqrt
clr	robCompositions::cenLR
ilr	robCompositions::pivotCoord
Auto	base::scale(x, center=TRUE, scale=apply(x,2,sd))
Pareto	base::scale(x, center=TRUE, scale=apply(x,2,sqrt(sd)))
Range	base::scale(x, center=TRUE, scale=apply(x,2,max)-apply(x,2,min))
Level	base::scale(x, center=TRUE, scale=apply(x,2,mean))
Vast	base::scale(x, center=TRUE, scale=apply(x,2,var)/(apply(x,2,mean)))
VSN	+vsn::vsn2

Table 2.1: Pre-treatment methods and examples of their R functions: The packages marked with * are available and can be downloaded at `https://github.com/walachja/pretreatment`. The packages with symbol + are part of the Bioconductor project (Gentleman et al., 2004) and they need a special installation.

2.4.1 Data sets

As an example for omics disciplines, the focus is given to data from Metalobomics. Altogether, three data sets are considered in the study. The first data set MTBLS17 is used only for classification, the second and third data set MTBLS59 and MCAD is used for both classification and feature selection.

The data set MTBLS17 (Ressom et al., 2012) is freely available from the MetaboLights repository `https://www.ebi.ac.ukasMTBLS17`, and it originates from blood samples of patients with liver cirhosis. Hepatocallullar carcinoma (HCC) is the fifth most common cancer type and 80% of HCC is developed from liver cirhosis. Only the positive ion part (ESI-) is considered. There are 184 control patients with liver cirhosis and 78 patients with HCC disease. The samples were analyzed in a non-targeted setting using ultra performance liquid chromatography coupled with a hybrid quadrupole time-of-flight mass spectrometry (UPLC-QTOF MS). The pre-processing of the data is described in Ressom et al. (2012). Due to subsequent resampling, the final dimensionality of the

data set is 372 samples with 941 features.

The second data set used here is again open-access data, MTBLS59 (Franceschi et al., 2012; Wehrens et al., 2011) from the MetaboLights repository `https://www.ebi.ac.uk` as MTBLS59. In total, twenty apple samples were analyzed by Liquid chromatography–mass spectrometry (LC-MS) in the study. Ten out of those were spiked with naturally occurring compounds, so the data are separated into two groups. An advantage of spiking the compounds into samples is that the features capable to distinguish the groups are known. In metabolomics such features are referred to as biomarkers. The data were pre-processed as in Wehrens et al. (2011) where only the first nine mintes of the chromatography were subtracted. Thus, the final data size is 20 samples and 197 features. The number of true biomarkers is 5, which corresponds to around 2.5% of all features.

The last data set MCAD (Najdekr et al., 2015) is based on plasma samples collected from 8 healthy newborn babies with Medium chain acyl-CoA dehydrogenase deficiency (MCADD OMIM# 201450). It is a disease called fatty acid oxidation disorder (FAODs). As a control group, plasma of 25 newborns were used. The LC-MS untargeted analysis was performed, and due to feature reduction techniques and subsequent sampling from the disease group the final data has 50 samples and 279 features. Even though the true biomarkers are not certainly known, previous analyses of the data suggest 5 features to be most likely the biomarkers. The data can be found in the R package *robCompositions* (Templ et al., 2017).

2.4.2 Zero imputation

The essential requirement to work with log or log-ratio transformations is to have data values which are all strictly positive. Since the measurements in these data sets will not be negative, this means that zero values need to be avoided and replaced by positive values. Firstly, one should keep in mind that there are several reasons why there are zero values in the data. It might be simply because the measured value is not present at all. This zero type is called essential zero. However, in metabolomics, transcriptomics and other omics fields the zeros appear as values below detection limit, so-called rounded zeros. Another possibility, as for example in metabolomics, is that rounded zeros usually arise as a pre-treatment step: if the value is below a certain threshold it is suppressed to zero because of the inaccuracy of the measurement device. These values are sometimes also called below detection values.

There are several methods available to dealing with zero imputation. Some of them are closely related to methods for missing values imputation, when zeros are viewed as

missing values. The easiest method is to replace all the zeros with 2/3 of the detection limit of the measurement device. If the detection limit is unknown, 2/3 of the lowest value of the variable under consideration can be used instead. However, if many zeros are present, they are all replaced by the same value, thus lowering the variability of the data set.

In the compositional data analysis (CoDa) context, there are several methods (e.g. Templ et al. (2016); Martín-Fernández et al. (2012, 2015)) for the imputation of zero values, based e.g. on the covariance structure of the data. For example, the algorithm proposed in Templ et al. (2016) can be described as follows. Firstly, the data are transformed to a specific ilr transformation (pivot coordinates). Secondly, Tobit regression (Scott Long, 1997) is applied and the zeros are replaced by the expected values. Thirdly, the data are transformed back to the original space. Lastly, the whole procedure is iteratively repeated until the imputed values stabilize. The algorithm can be applied from the R package *robCompositions* (Templ et al., 2017) as the function *imputeBDLs()*.

2.4.3 Classification and feature selection

Classification analysis is a frequent task in omics disciplines. Creating a model for the separation of the samples into groups is important for two reasons. Firstly, for prediction of the newly incoming observations, and secondly, for the description of the model. Here the focus is given to the first task – prediction of class membership for a new observation. Below we only consider two-group classification problems. Another common goal in omics disciplines is to interpret the model in terms of identifying important features. In the classification context, the important variables are those which help separating the groups in the data.

Classification and feature selection method

There are many methods for feature selection – one of the simplest yet frequently applied methods for a two-group problem is a univariate two-sample t-test for each candidate variable. Here we will make use of a robust multivariate method for classification, the Partial robust M discriminant analysis classifier (PRM-DA) (Hoffmann et al., 2016), followed by a score for feature evaluation. PRM-DA is a robust version of the Partial least-squares discriminant analysis classifier (PLS-DA) (Wold et al., 2001; Pérez-Enciso and Tenenhaus, 2003), a frequently used multidimensional classification method. The PRM-DA, as well as PLS-DA allow to express the possibly high-dimensional data information in a low-dimensional space where discriminant analysis is carried out. The methods can

deal with multi-collinearity, i.e. with highly correlated predictor variables, and with a situation where the number of samples n can be much smaller than the number of variables d. The PRM-DA method, which is highly robust against data outliers (and even robust against mis-labeling) consists of two steps.

Firstly, Partial robust M regression (PRM) (Serneels et al., 2005), a robust version of Partial least-squares (PLS) regression, with binary response variable is computed. In order to simplify the notation, from now on column vectors are considered. The response vector $\mathbf{y}$ of length n carries the group information coded as -1 and 1 for two groups, say, A and B. The predictor variables are stored in the $n \times d$ matrix $\mathbf{X}$ matrix. The objective of the method is to find directions $\mathbf{a}_h$ with

$$\mathbf{a}_h = \underset{\mathbf{a}}{\operatorname{argmax}} \operatorname{cov}^2 (\mathbf{Xa}, \mathbf{y}) , \tag{2.9}$$

for $h \in \{1, \ldots, H\}$ subject to $||\mathbf{a}_h|| = 1$ and $\mathbf{a}_h^T \mathbf{X}^T \mathbf{X} \mathbf{a}_i = 0$ for $1 \leq i \leq h$. The directions $\mathbf{a}_h$ are stored in the columns of the matrix $\mathbf{A}$, and the scores $\mathbf{T}$ are defined as $\mathbf{T} = \mathbf{XA}$.

"cov" in Equation (2.9) stands for the covariance. While PLS uses the classical sample covariance estimation, PRM employs a robust covariance estimator by assigning weights to the observations. The weights correspond to the outlyingness of the observations and they are iteratively updated during the estimation procedure.

The second step is a linear discriminant analysis (LDA), carried out for the scores $\mathbf{T}$. The rows of $\mathbf{T}$, denoted by $\mathbf{t}_i$, for $i = 1, \ldots, n$, are assigned to that group $k \in \{A, B\}$, for which the discriminant score δ_k is the highest, with

$$\delta_k = \mathbf{t}_i^T \hat{\mathbf{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\mathbf{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k + \log (\pi_k) , \tag{2.10}$$

with the group prior probabilities π_k. While for PLS-DA the classical estimates for $\hat{\boldsymbol{\mu}}_k$ and $\hat{\mathbf{\Sigma}}$ are taken, PRM-DA uses the observation weights w_i obtained from the first step (Todorov and Pires, 2007),

$$\hat{\mathbf{\Sigma}} = \frac{1}{(\sum_{i=1}^{n} w_i) - 2} \sum_{k \in \{A,B\}} \sum_{i \in C_k} w_i (\mathbf{t}_i - \hat{\boldsymbol{\mu}}_k) (\mathbf{t}_i - \hat{\boldsymbol{\mu}}_k)^T , \tag{2.11}$$

and

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i \in C_k} w_i \mathbf{t}_i}{\sum_{i \in C_k} w_i} \quad \text{for} \quad k \in \{A, B\} ,$$

where C_k are the indexes of observations belonging to group k.

The optimal number of components H for PRM-DA is determined by K-fold cross-validation (CV). For all considered numbers of components and for each fold, the mean weighed misclassification rate (mwmcr) is computed. The number of components with the lowest average mwmcr is selected as optimal.

PRM-DA is implemented in the R package *sprm* Serneels. and Hoffmann (2016) as the function *prmda* and *prmdaCV*.

For the feature selection problem the Variable Importance in Projection (VIP) (Wold et al., 1993; Favilla et al., 2013) is employed, which is a commonly used method for the evaluation of the importance of individual explanatory variables, mainly in the context of projections to a lower dimensional space. Thus, it can also be used in combination with PRM-DA. The VIP summarizes the contribution of a variable to the model, and is defined for the jth variable as

$$\mathrm{VIP}_j = \sqrt{\frac{d}{\sum_{h=1}^{H} R^2\left(\mathbf{y}, \mathbf{t}_h\right)} \sum_{h=1}^{H} a_{hj}^2 R^2\left(\mathbf{y}, \mathbf{t}_h\right)}, \tag{2.12}$$

where d is the number of variables, H is the number of PRM components, a_{hj} denotes the jth component of the loadings vector $\mathbf{a}_h$, and $R^2\left(\mathbf{y}, \mathbf{t}_h\right)$ is the fraction of variance in $\mathbf{y}$ explained by the hth PRM component. A typical rule is that those variables with index j are selected as important for which $\mathrm{VIP}_j > 1$ (Chong and Jun, 2005; Lazraq et al., 2003). Alternatively, one can select the l variables with the highest VIP scores as the most important ones.

2.4.4 Application and evaluation

In order to apply and evaluate the different normalization, transformation and scaling methods, the procedure was done as follows:

1. Zero replacement. All three data sets contain zero values, which means that several methods (i.e. Contrast, cLOESS, Log and clr) cannot be applied directly. Thus, as suggested in Section 2.4.2, zeros were replaced by strictly positive values. For simplicity, and since a detection limit was not available, we replaced the zeros by 2/3 of the smallest positive value of the corresponding variable.

2. Normalization methods. All 17 normalization, transformation and scaling methods were applied to the original data sets. Only if a method cannot deal with zeros, the data set with the replaced zeros is used.

3. Two-step normalization methods. All 17 pre-treatment methods were combined with log-transformation or with autoscaling. The logarithm is often used to deal with heteroscedasticity. Autoscaling is frequently used in multidimensional methods such as PCA. Both, log-transformation and autoscaling can be applied either before or after another pre-treatment method. However, there are several exceptions, since e.g. the clr transformation produces negative values, the use of the logarithm afterwards is not

possible (and would also not make sense). In total, 35 two-step normalization methods are considered. Together with the 17 "one-step" methods and with the raw untreated data, this gives 53 methods in total.

4. Optimal number of components of PRM-DA. As outlined in Section 2.4.3, 10-fold CV was performed for PRM-DA to estimate the optimal number H of components. This was done separately for each of the 53 methods.

5. Evaluation. 2/3 of all samples were randomly selected to estimate a PRM-DA model with the optimal number of components computed in the previous step. The remaining third of the samples were treated as test data and their group membership is predicted. The whole procedure is repeated 100 times.

The predicted groups memberships are compared with the true group labels in terms of the *Accuracy*, which is the proportion of correctly predicted groups labels in the corresponding test set. Note that since 100 replications are carried out, and since this is done for each of the three data sets, in total 300 accuracy values are computed for each of the 53 pre-treatment methods, which are summarized by their average.

For feature selection, the VIP scores defined in Section 2.4.3 were used. In each replication, the indices of the variables with the 10 highest VIP values were extracted. Since the true important feature structure is known for two of the examined data sets (5 in both cases), the number of true biomarkers among these 10 highest is computed, leading to 200 numbers, which are averaged separately for each pre-treatment method.

Since both, accuracy in classification and in feature selection may be important, the results are combined in order to provide a better overview. This is done by ranking the different pre-treatment method separately according to the resulting classification and feature selection accuracy, with rank 1 being the winner. The *average of the two ranks* is used as an overall evaluation of the methods.

2.4.5 Results

Figures 2.2 and 2.4 summarize all results of the analyses. The rows in the plots refer to the 17 "one-step" pre-treatment methods and the raw untreated data. If possible, each method can be extended to a two-step method with the logarithm or with autoscaling, and both can be applied beforehand or afterwards. For that reason, there are several different plot symbols: black dots represent the 18 "basic" methods (including untreated raw data), "L" represents combinations where the logarithm is applied first, "l" are combinations where the logarithm is applied afterwards, and similarly "A" and "a" for combinations with autoscaling. Several of the "two-step" methods cannot be applied or they are not meaningful to be carried out. For example, since clr already uses the

logarithm, it is useless and even impossible (due to negative values) to apply the logarithm after or before this transformation.

Accuracy of class prediction. Figure 2.2(a) shows the effects of the pre-treatment methods sorted by highest classification accuracy (of the "basic" methods). The four best performing methods are based on (or at least include) logarithms of the original values: clr, Cyclic LOESS, VSN and Log. Considering also two-step methods, it can be seen that the logarithm (as a first or second step) improves the accuracy of prediction. This suggests that heteroscedasticity or strong skewness is present in the analyzed data sets which is improved by using the logarithm. The next best four methods are scaling methods: Level, Range, Auto and Pareto scaling. The performance of these methods is improved by first log-transforming the data. Although the errors can be increased by these scaling methods, the variables important for group separation can be in lower abundances. Note that the performance PRM-DA is dependent on the scale of the variables, because it looks for highest covariance between predictors and response. Vast scaling, which uses the coefficient of variation as a scaling factor, is the only scaling method with severely poorer performance. The coefficient of variation can become very large for low abundances, which can lead to an inflation of errors. The next seven methods (PQN, cSplines, TSN, Sqrt, lBase, Quantile, and also Vast) have similar performance as for the raw data without any pre-treatment. Non-linear baseline normalization (nlBase) and Contrast normalization achieved the worst performance. This might be based on a relatively complex estimation procedure, and possibly also on overfitting (smoothing splines need to be fitted in case of nlBase).

Feature selection results. Figure 2.2(b) shows the results for the feature selection accuracy, and the methods (rows in the figure) are sorted now according to this accuracy. The best performing methods are Cubic Splines and Quantile normalization. The methods are similar, since both try to force all features to have similar distributions. Surprisingly, using the raw data without pre-treatment also leads to a top performance. This means that choosing an improper pre-treatment method can severely lower the performance of feature selection analysis. Almost all two-step methods decrease the performance of the "basic" methods, with the exception of cLOESS, VSN, and three scaling methods.

The comparison of the methods concerning the correct identification of the 5 biomarkers is shown in more detail in Figure 2.3, separately for the two data sets. The gray scale shows how many times out of the 100 replications the truly important variables (in the columns) have been identified. White color means that the corresponding variable was never among the ten variables with highest VIP scores, whereas black color shows perfect results, correct identification in each replication. In the rows are the "basic"

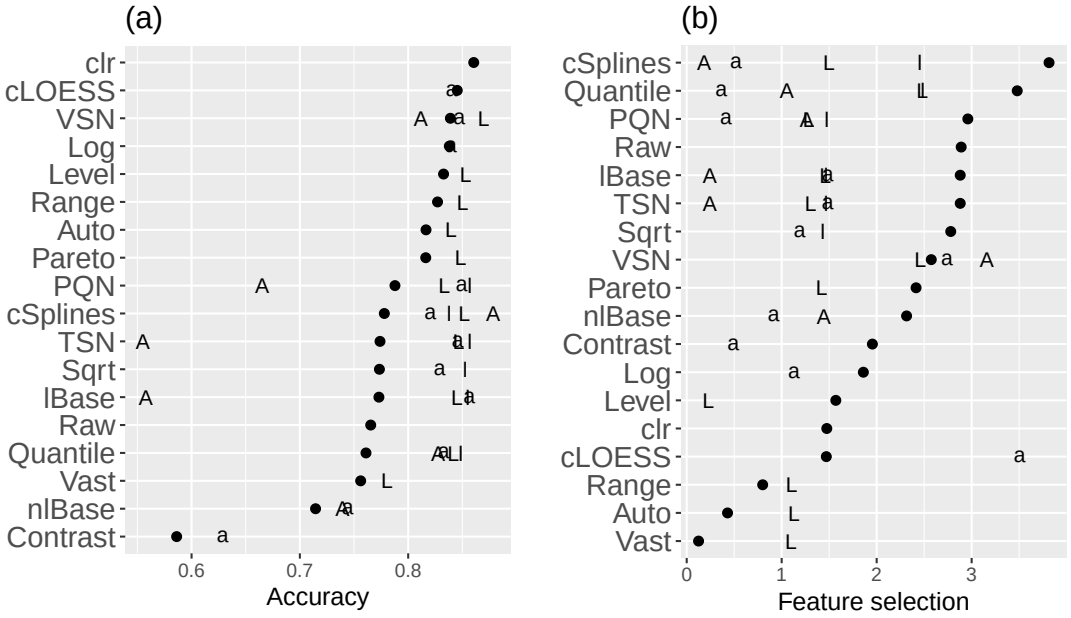


Figure 2.2: Sorted results of (a) accuracy and (b) feature selection of different methods: black dot refers to the result of the "basic" method written on the left side; "A" corresponds to first applying autoscaling and afterwards the "basic" method; "a" means that autoscaling is applied after the "basic" method; similar for "L" and "l" with the logarithm instead of autoscaling.

pre-treatment methods or their two-step variations if they identified a bigger number of truly important features. Figure 2.3(a) shows the results for the MCAD and Figure 2.3(b) for the MTBLS59 data set.

Many methods correctly identified the first two truly important features in the MCAD data set, and the last three in the MTBLS59 data set. The best performing methods identified in Figure 2.2(b), cSplines and Quantile normalization, give also a very similar answer here concerning the biomarker identification. It seems, however, that it is difficult for the methods to identify correctly all 5 biomarkers in the two data sets.

Figure 2.4 summarizes again the results, but the methods are ordered now according to the average of the ranks among the methods for the classification and feature selection task, see plot (a). Plot (b) presents the average proportion of samples that received weights from PRM-DA which are lower than 0.1, i.e. the proportion of identified "outliers".

According to Figure 2.4(a), the clr transformation is the best performing method if we consider only "basic" (one-step) pre-treatment methods, even though the feature selection results are not very good. Clr is followed closely by VSN. VSN may have achieved superior results because it combines variance stabilization with normalization

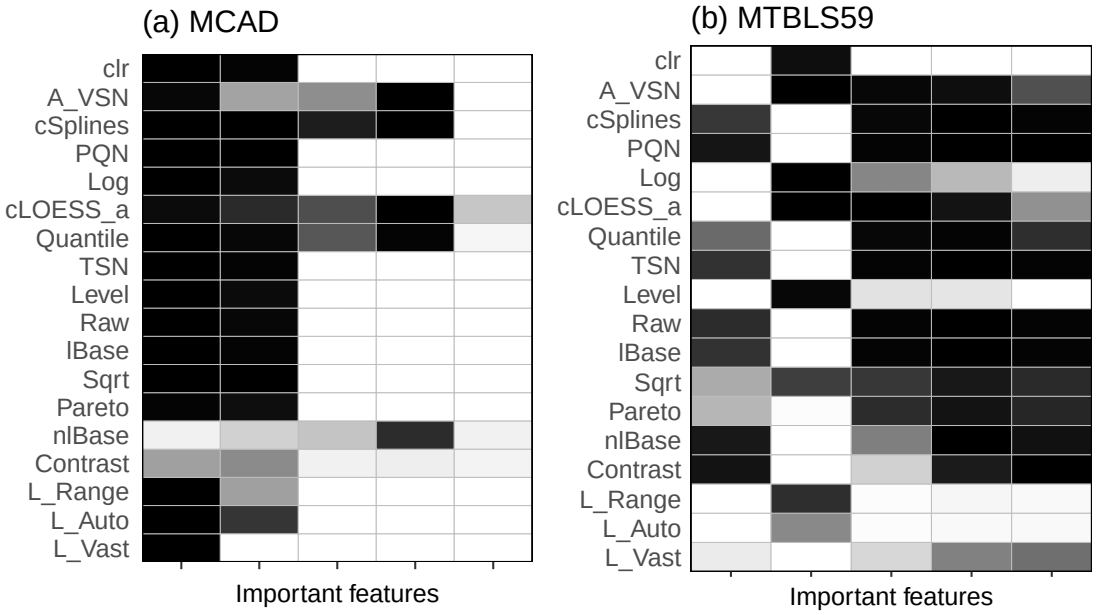


Figure 2.3: Performance of the different methods (rows) in identifying the true important features (columns): gray scale represents how often (out of 100 replications) the biomarker was identified correctly within the 10 top VIP variables (black=always, white=never). Plot (a) shows the results for MCAD, plot (b) for MTBLS59.

between the samples. Furthermore, there is a group of moderately performing methods (cSplines, PQN, Log, cLOESS, Quantile, TSN, Level, lBase, Sqrt and Pareto), which have achieved a similar average rank. The five least performing methods (nlBase, Contrast, Range, Auto, Vast) have noticeably lower ranks than the other methods. If two-step methods are considered, then the performance improves in many cases if the logarithm is applied. On the other hand, the use of autoscaling leads for most methods to a decrease of the performance. The reason may be that errors (biological, technical or others) of features with small abundance and thus presumably small variances are increased.

The results in Figure 2.4(b) show a trend that the least ranked methods lead to a somewhat higher proportion of outliers for PRM-DA. This means that those methods differs more from multivariate normality which can be one reason for a poorer performance. To demonstrate this fact more clearly, Figure 2.5 shows PRM-DA score plots of one of the replications (random training sample) for a poor (Contrast) and well (VSN) performing method for the biggest data set MTBLS17. After Contrast normalization, the data in both groups are highly skewed, which leads to many outliers in the PRM estimation procedure, and consequently to a low classification accuracy. VSN leads to elliptically

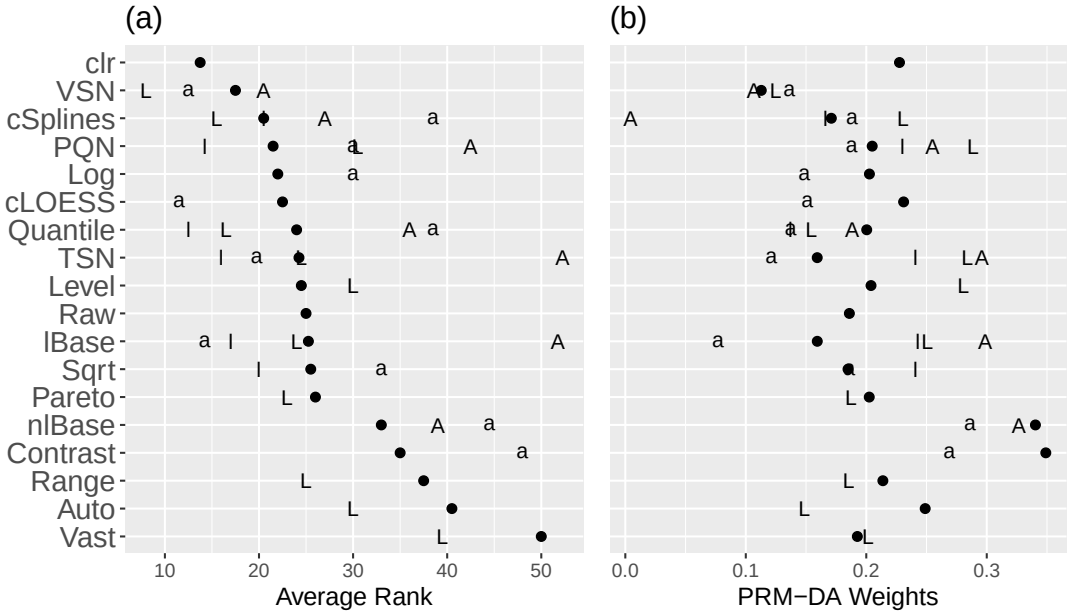


Figure 2.4: Sorted results of (a) average rank and (b) proportion of small weights from PRM-DA: black dot refers to the result of the "basic" method written on the left side; "A" corresponds to first applying autoscaling and afterwards the "basic" method; "a" means that autoscaling is applied after the "basic" method; similar for "L" and "l" with the logarithm instead of autoscaling.

distributed scores in both groups, which is preferable for the subsequent robust LDA.

The methods behave differently depending on the purpose of the analysis. For example, clr has superior behavior for the accuracy of the prediction but not very good for feature selection behavior. This is similar for some two-step methods. For example, applying autoscaling as a second step after PQN improves prediction accuracy but worsens feature selection performance. The same applies to cSplines after autoscaling. The cLOESS in combination with autoscaling behaves in an opposite way, since it improves feature selection performance but worsens the accuracy of the prediction. Thus, even though it is a common practice to use two-step methods, one should keep in mind the purpose of the analysis and choose appropriate combinations of the used methods.

2.5 Discussion and conclusions

The choice of the appropriate pre-treatment method depends on the biological tasks and questions, on the properties of the data, and of course also on the data themselves. The selection of a proper pre-treatment method is a crucial step in the analysis, since it can

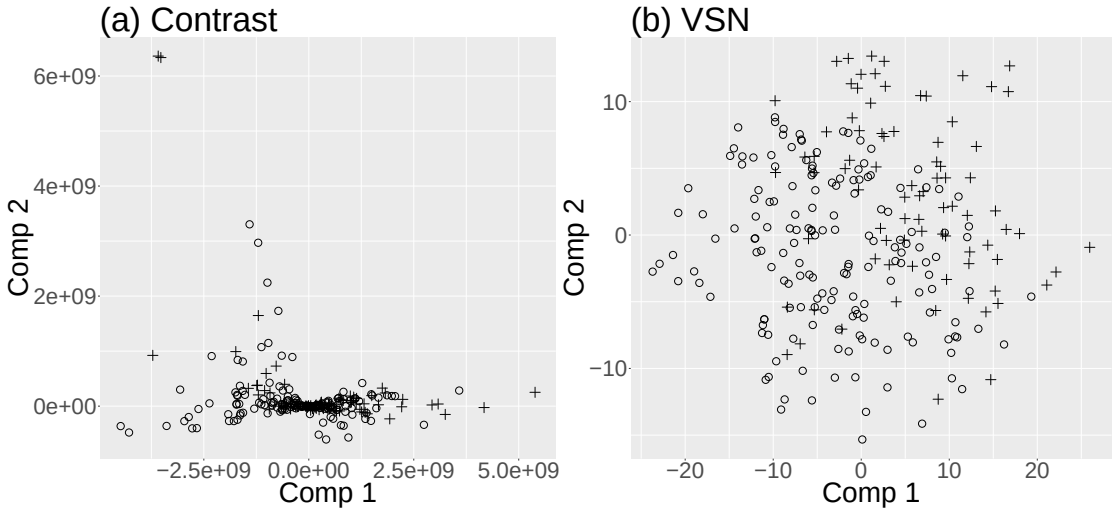


Figure 2.5: Score plots from PRM-DA for one specific training set of the MTBLS17 data set, based on (a) Contrast normalization and (b) VSN.

severely improve, but also deteriorate the results of the analysis.

In this article several normalization, transformation and scaling methods were discussed. These pre-treatment methods are frequently used in practice, but we admit that this is not an exhaustive list of methods. Moreover, in specific fields there might be very specific methods which are adequate for the specific application.

Another limitation is the type of evaluation demonstrated here for three different data sets. Although the dimensions of the data are quite different, the data sets are generated via mass spectrometry, and for more general conclusions also other types of data might have to be considered. Moreover, the type of the analysis with PRM-DA and the VIP measure is specific, and other forms of evaluations might lead to different conclusions. We also evaluated and compared the results with the Selectivity Ratio (Rajalahti et al., 2009a,b) as an alternative to the VIP measure. Overall, these results were similar to those reported here, but biomarker identification based on the VIP measure turned out to be more consistent. In order to be a bit more general, some conclusions from analyses in other papers are briefly summarized.

The study Kohl et al. (2012) focuses on NMR metabolomics data and concludes that for classification analysis the best pre-treatment methods are Quantile and Cubic Spline normalization, and VSN. The paper Hochrein et al. (2015) came with the same conclusion but also adds PQN to the best methods. The work Li et al. (2016) examines MS data and conclude that Log, VSN and PQN are among the best methods to use for classification and feature selection. In Gromski et al. (2015), the author combines NMR

and MS data and focuses only on scaling methods. His findings suggest that Vast scaling performs best, which is not in line with the findings in our study. The paper Saccenti (2017) lists the worst performing methods, including Non-linear baseline normalization, Cyclic Loess and Contrast normalization. However, one should keep in mind that these studies did not consider all the pre-treatment methods (one-step and two-step) which have been analyzed here.

For example, the clr transformation was not analyzed in any of these studies. In Filzmoser and Walczak (2014), clr was compared to PQN on simulated data for the accuracy in feature identification, and PQN turned out to be clearly preferable. However, also the source information for the clr transformation, the pairwise log-ratios were employed in this work, which are very competitive to PQN. Recently, Walach et al. (2017) developed a method which also makes use of pairwise log-ratios, but more efficiently, and the method is also robust against data outliers. We applied this method to the data sets in this study, and it identifies on average 3.96 of the true important features. This is better than the best pre-treatment method cSplines evaluated here, which had an average of 3.82.

Based on the analyses performed in this study, and taking into account other studies listed above, several recommendations can be provided. Firstly, it is crucial to apply a pre-treatment method. Secondly, the pre-treatment method strongly depends on the goal of the analysis. For classification analysis the methods based on logarithms (clr, Cyclic Loess and VSN) perform best. Even the use of the logarithm after a basic normalization method can improve the results. Considering feature selection analysis, Cubic Splines and Quantile normalization are recommended. In practice, there might still be further criteria for the selection of pre-treatment methods, such as the simplicity of the method, or the availability in standard software packages. Moreover, sometimes the data are reported with zeros, which excludes methods that are based on the logarithm. The logarithm may also produce negative values, which are not adequate for some purposes. In any case, one must be aware that if a pre-treatment is employed, the interpretation for the selected feature may change – because of the pre-treatment employed.

Acknowledgments

This research is supported by the Austrian Science Fund (FWF) and Czech Science Fund (GACR), project number I 1910-N26 (15-34613L). The authors would like to thank an anonymous reviewer for very valuable comments.

CHAPTER 3

Robust biomarker identification in a two-class problem based on pairwise log-ratios

A new method, robust Pair-wise Log-Ratios (rPLR), is proposed for the identification of biomarkers, distinguishing between two groups of observations. The method can cope with the size effect problem, since it is based on log-ratios between the values of all pairs of variables. rPLR makes use of the variance of pairwise log-ratios, computed for the single groups and for all data jointly. When using a robust estimator of variance (or scale), the method is highly robust against data outliers. The robustness weights are aggregated and displayed in a diagnostics plot, which allows to reveal outlying cells in the data matrix.

3.1 Introduction

"Omics" approaches (e.g. genomics, proteomics, metabolomics) are important platforms for interpreting and understanding complex biological systems. Nowadays, the use of different types of hyphenated techniques such as e.g., LC-MS, UPLC-MS, are standard and there is a need for methods being capable of dealing with the data coming from this field. This paper proposes a robust method based on Pair-wise Log-Ratios (rPLR) for the identification of the key features, which are able to distinguish between two groups of

samples (e.g. patients with and without a certain disease) (Monteiro et al., 2013; Lindon et al., 2003). In this context, this problem is known as biomarker identification. Here, we will focus on a situation when the so-called size effect is present in the data. The term size effect refers to measured samples which have different sample concentrations. The size effect is obviously undesirable, and it occurs if the true signal cannot be directly observed. Instead, the true signal multiplied by a constant is measured. The constant is in general different for each sample which is the basic problem with the size effect. A typical example of the size effect is the analysis of urine samples.

There are several possibilities how to deal with the size effect. A standard procedure is preprocessing of the data by applying certain normalizations or transformations. A widely used normalization method is total sum normalization (TSN), where the values of each sample are divided by their sum. Thus, after TSN, the values of each sample sum up to one. However, for the purpose of biomarker identification, TSN is problematic since it can mask the biomarkers (Filzmoser and Walczak, 2014).

An alternative is probabilistic quotient normalization (PQN) (Dieterle et al., 2006). Let us assume an $(n \times d)$ data matrix $\mathbf{X}$, with n samples and d measurements, and with the matrix elements x_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. For a sample $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$, PQN estimates the scaling constant s_i as the median of the ratios of the elements of $\mathbf{x}_i$ to "reference" values $x_{\text{ref},j}$ for each variable, $s_i = \text{median}(x_{i1}/x_{\text{ref},1}, \ldots, x_{id}/x_{\text{ref},d})$. The reference values are the column medians or means of $\mathbf{X}$ (Dieterle et al., 2006). The normalized values of the ith sample are

$$\mathbf{x}_i^{PQN} = \left(\frac{x_{i1}}{s_i}, \ldots, \frac{x_{id}}{s_i} \right),$$

for $i = 1, \ldots, n$. PQN assumes that the majority of the variables is not different between the analyzed groups.

In the paper Filzmoser and Walczak (2014), several normalization and transformation methods were examined for a subsequent identification of biomarkers. Besides TSN and PQN, also transformations from compositional data analysis, as well as pairwise log-ratios were investigated (Pawlowsky-Glahn and Buccianti, 2011). It turned out that PQN was the most preferable normalization method for size effect removal in the context of biomarker identification. Good results could also be achieved with the pairwise log-ratio approach, but since the number of distinct variable pairs is $d(d-1)/2$, this method becomes impracticable in case of high-dimensional data, but also the results cannot be easily interpreted.

In principle, the size effect problem can be solved by working with ratios rather than with the original information. This can be easily shown by assuming that the

true signal information is $\mathbf{x} = (x_1, \ldots, x_d)$. In presence of a scaling constant we observe $s \cdot \mathbf{x} = (s \cdot x_1, \ldots, s \cdot x_d)$. However, the ratios between any two variables of the true signal, x_j/x_k, carries the same information as the corresponding ratios of $s \cdot \mathbf{x}$, since $(s \cdot x_j)/(s \cdot x_k) = x_j/x_k$. Thus, the relevant information is contained in the ratios between the variables.

As noted in Aitchison and Shen (1980), ratios are not easy to deal with, because their variances are non-symmetrical, since $\text{var}(x_j/x_k) \neq \text{var}(x_k/x_j)$. This was solved by using logarithms of ratios, so called log-ratios, which meet the property of symmetry, since $\text{var}(\ln(x_j/x_k)) = \text{var}(\ln(x_k/x_j))$. Log-ratios are used in the field of compositional data analysis (Pawlowsky-Glahn and Buccianti, 2011).

The main goal of this study is to present a new method for biomarker identification based on robust Pair-wise Log-Ratios: rPLR (Section 3.2) and to examine its behavior. The results of rPLR are compared with other normalization methods. Another focus in this paper is robustness. Robust statistical methods are often used since they can generally deal with data where outliers are present, see, for example Liang and Kvalheim (1996); Liang and Fang (1996). Since most real-world measurements – including "omics" data – contain outliers, robust procedures are preferable. The proposed method is straightforward to robustify, and thus its robustness properties are examined in simulation studies in Section 3.3. Section 3.4 presents new ways of outlier diagnostics, which also lead to interesting findings in a real data example in Section 3.5. The final Section 3.6 provides concluding remarks.

3.2 Method rPLR

Consider an $n \times d$ data matrix $\mathbf{X}$, where the observations originate from two groups. Let $\mathbf{X}^{(1)}$ denote the sub-matrix with the n_1 observations in the rows from the first group, and $\mathbf{X}^{(2)}$ the corresponding matrix with n_2 observations of the second group, and $n_1 + n_2 = n$. The matrix elements of $\mathbf{X}^{(l)}$ are denoted by $x_{ij}^{(l)}$, for $i = 1, \ldots, n_l$, $j = 1, \ldots, d$, and $l = 1, 2$.

3.2.1 Variation matrix

The proposed method builds on the variation matrix $\mathbf{T}$ (Aitchison, 1986; Pawlowsky-Glahn et al., 2015), with the elements t_{jk} defined as:

$$t_{jk} = \text{var}\left[\ln\left(\frac{x_{1j}}{x_{1k}}\right), \ln\left(\frac{x_{2j}}{x_{2k}}\right), \ldots, \ln\left(\frac{x_{nj}}{x_{nk}}\right)\right], \tag{3.1}$$

where $j, k = 1, \ldots, d$, and "var" denotes the variance. The elements of the variation matrix report the variability of the log-ratio of a pair of variables. The smaller the value of t_{jk} is, the more the log-ratio tends to be a constant. In this case, the corresponding variables can be considered as being proportional. The variation matrix $\mathbf{T}$ is symmetric (see Section 3.1), and the diagonal elements are zero.

Besides the variation matrix $\mathbf{T}$ based on all observations jointly, the individual group variation matrices are considered as well. Let us denote $\mathbf{T}^{(l)}$ as the variation matrix of group l, for $l = 1, 2$, with the elements defined as

$$t_{jk}^{(l)} = \text{var}\left[\ln\left(\frac{x_{1j}^{(l)}}{x_{1k}^{(l)}}\right), \ln\left(\frac{x_{2j}^{(l)}}{x_{2k}^{(l)}}\right), \ldots, \ln\left(\frac{x_{n_l j}^{(l)}}{x_{n_l k}^{(l)}}\right)\right], \tag{3.2}$$

for $j, k = 1, \ldots, d$. Thus, the variation matrices of the individual groups consider only the observations from their own groups.

3.2.2 Test statistic

For biomarker identification, the following statistic V_j is proposed,

$$V_j = \sum_{k=1}^{d} \frac{n_1 \cdot \sqrt{t_{jk}^{(1)}} + n_2 \cdot \sqrt{t_{jk}^{(2)}}}{(n_1 + n_2) \cdot \sqrt{t_{jk}}}, \qquad \text{for } j = 1, \ldots, d. \tag{3.3}$$

If the jth variable is not a biomarker, the jth column (and row) of all three sources of information $\mathbf{T}$, $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$ will have similar structure. For this reason, each term of the sum in (3.3) will be approximately around one for all non-biomarkers k. On the other hand, if the jth variable is a biomarker, $t_{jk}^{(1)}$ and $t_{jk}^{(2)}$ will be different, and tentatively much smaller than t_{jk}, for all k. The resulting V_j will then be considerably smaller than for non-biomarkers. So, the smaller the value of the statistic (3.3) is, the less similar the groups are with respect to this jth variable.

Note that in Equation (3.3), the elements of the variation matrix are weighted with the number of samples of both groups. In case of equal sample sizes (balanced setting) it is easy to see that V_j can be simplified to

$$V_j = \sum_{k=1}^{d} \frac{\sqrt{t_{jk}^{(1)}} + \sqrt{t_{jk}^{(2)}}}{2 \cdot \sqrt{t_{jk}}}, \qquad \text{for } j = 1, \ldots, d. \tag{3.4}$$

Since the distribution of V_j is not known, it is not straightforward to define a cut-off value which would allow to distinguish between biomarker and non-biomarker.

For "omics" data, however, one could argue that the vast majority of variables is independent, with a similar distribution. Since d is usually big, the central limit theorem

would then imply normal distribution, at least for those V_j referring to non-biomarkers (so, for the vast majority). Although normality cannot be proven formally, our simulation study shows that the values V_j follow approximately a normal distribution. The square root in the statistics 3.3 and 3.4 is used in order keep the values of V_j more symmetric, hence closer to normality.

We consider a normalized version

$$V_j^* = -\frac{V_j - \bar{V}}{s_V}, \qquad \text{for } j = 1, \ldots, d, \tag{3.5}$$

with the arithmetic mean

$$\bar{V} = \frac{1}{d} \sum_{k=1}^{d} V_k$$

and the empirical standard deviation

$$s_V = \sqrt{\frac{1}{d-1} \sum_{k=1}^{d} (V_k - \bar{V})^2}.$$

Because of the minus sign in (3.5), now big values of V_j^* point are potential biomarkers, which is easier to grasp in a visual presentation of the outcome. Following the argumentation from above, most values V_j^* will be approximately standard normally distributed, and we will use the standard normal quantile $u_{0.975} \approx 1.96$ as cut-off for biomarker identification. In other words, all variables with index j, where $j \in \{1, \ldots, d\}$, are identified as biomarkers, if their statistic $V_j^* > u_{0.975}$. Note that the statistic V_j^* is based on all bivariate information with the jth variable, and also the grouping information is considered.

Although this approach using approximate normality was very useful in our experiments, one could also employ randomization tests (e.g. Kempthorne (1952); Edgington and Onghena (2007)) as an alternative. Randomization tests do not assume normality or any other distribution of the data, but they are computationally much more demanding.

3.2.3 Estimation of the standard deviation

The performance of the rPLR method crucially depends on how the involved variation matrices are estimated. More clearly, it is important which estimator of variance is used for "var" in Equations (3.1) and (3.2). In the following we discuss different possibilities to estimate its square-root, the standard deviation. The standard choice would be the empirical standard deviation. However, in presence of outliers it is well known that this classical estimator in not robust and thus can yield heavily biased results (Yohai and

Zamar, 1988). As a consequence, (some of) the values V_j would be spoiled, and biomarker identification based on V_j^* would become unreliable.

Fortunately, the proposed method can be easily robustified by employing a robust estimator of scale. We consider the following two options.

Median absolute deviation

The Median Absolute Deviation (MAD) is probably the most common robust estimator of standard deviation. For a univariate sample $\mathbf{y} = (y_1, \ldots, y_n)$, it is defined as

$$\mathrm{MAD}(\mathbf{y}) = 1.48 \cdot \mathrm{median}_i \left| y_i - \mathrm{median}(\mathbf{y}) \right|. \tag{3.6}$$

The MAD is highly robust since it can resist against up to 50% outliers. A disadvantage of this estimator is the low statistical efficiency of around 37% for normally distributed data (Rousseeuw and Croux, 1993). The statistical efficiency of an estimator refers to its precision, and it can be described as the number of observations needed to achieve a given performance (Lambeth et al., 1983). Since the classical standard deviation obtains the highest possible efficiency (100%) under normality, one would need 63% more observations to achieve the same performance with the MAD as with the standard deviation.

τ-estimator of scale

The τ-estimator is a highly robust estimator of scale, but it also attains a high efficiency, tunable with two constants c_1 and c_2. This is particularly important when dealing with only few samples. It uses weights for the observations, defined as

$$w_i = \omega_{c_1}\left(\frac{y_i - \mathrm{median}(\mathbf{y})}{s_0}\right) \qquad \text{for } i = 1, \ldots, n, \tag{3.7}$$

with the weight function

$$\omega_{c_1}(u) = \left(1 - \left(\frac{u}{c_1}\right)^2\right)^2 I\left(|u| \le c_1\right) \text{ and } s_0 = \mathrm{MAD}(\mathbf{y}).$$

Then the τ estimator of scale is defined as

$$\sigma_\tau = \sqrt{\frac{s_0^2}{n}\sum_{i=1}^{n} \rho_{c_2}\left(\frac{y_i - \bar{y}_w}{s_0}\right)}, \tag{3.8}$$

where

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \quad \text{and} \quad \rho_{c_2}(u) = \min(c_2^2, u^2).$$

In order to combine good robustness properties with high efficiency, the recommended tuning parameters are $c_1 = 4.5$ and $c_2 = 3$. This leads to around 80% efficiency at normal distributions, while keeping the breakdown point at 50%, see Yohai and Zamar (1988); Maronna and Zamar (2002).

3.3 Simulation study

The goal of this section is to investigate the performance of rPLR under different scenarios, and to compare with other methods. In particular, we are interested in the robustness behavior, as well as in unbalanced settings, where the groups consist of different numbers of observations. Unbalanced data appear frequently in practice, since usually only a low number of patients suffering from a certain disease is available, while the control group may consist of much more persons. On the other hand, in some cases the situation can be exactly reverse, especially in the case when the procedure of extracting samples is invasive. Then there will be relatively many samples from patients and only few controls. For the proposed methods, however, it is irrelevant which group forms the minority.

3.3.1 Simulation design

For simulating the data, we use the close-to-reality setting as proposed in Filzmoser and Walczak (2014). However, we restrict ourselves only to the high-dimensional case ($d \gg n$), which we consider more relevant in this context. Accordingly, the columns $\mathbf{x}_j$ of a simulated data matrix $\mathbf{X}$ are generated as

$$x_j = N_j + (1 - S) \cdot [(c_j + a_j) \cdot r_j + B_j] \cdot e^{M_j}, \tag{3.9}$$

for $j = 1, \ldots, d$. Here, N_j represents the background noise and it is generated from a normal distribution $\mathcal{N}(0, 0.05^2)$. The size effect S is generated from $\mathcal{N}(0, 0.3^2)$, and it is independent of j. Further, $c_j = a_j / r_j$ stands for the component concentration, where a_j is the signal abundance and r_j the component absorptivity. These two parameters come from uniform distributions: $c_j \sim U[5, 10]$, $r_j \sim U[1, 10]$. The parameter a_j creates different abundance of the biomarker, and is defined as

$$a_j = \begin{cases} A & \text{for } j \in I_{bm} \text{ and observations of group 1} \\ 0 & \text{otherwise,} \end{cases}$$

where I_{bm} is an index set containing the indices of the biomarkers. Here, $d_0 = 20$ variables are simulated as biomarkers, and without loss of generality, they are arranged at the first 20 positions, $I_{bm} = \{1, 2, \ldots, 20\}$. The sign of A is alternated among subsequent

Table 3.1: Set of parameters for the simulation.

Setting	Biol. variance σ_B	Between class abundance A	Multipl. noise σ_M
1	+	+	+
2	+	+	−
3	+	−	+
4	−	+	+
5	−	−	−
6	−	−	+
7	−	+	−
8	+	−	−

where $\sigma_B(+) = 0.8$, $\sigma_B(-) = 0.2$, $A(+) = 1.8$, $A(-) = 1$, $\sigma_M(+) = 0.021$, $\sigma_M(-) = 0.007$.

variables. In the simulation study 480 non-informative variables (non-biomarkers) were created which leads to 500 variables in total. The biological noise is modeled by B_j generated from $\mathcal{N}(0, \sigma_B^2)$. Finally, the multiplicative noise is generated according to $M_j \sim \mathcal{N}(0, \sigma_M^2)$. The rest of the parameter values are listed in Table 3.1. The different parameter combinations result in eight different settings. The settings with high variance (σ_B^2 and/or σ_M^2) and at the same time low abundance (A), so settings number 3, 6, 8, will lead to poor signal-to-noise ratio. The opposite situations represent more clearer separation between the groups.

3.3.2 Simulation of outliers

In order to generate outliers, the standard deviation for the noise term was increased by a factor 10, thus $N_j \sim \mathcal{N}(0, 0.5^2)$. Also, the distribution of the multiplicative noise is modified to $M_j \sim \mathcal{N}(\pm 0.5, (10 \cdot \sigma_M)^2)$, where the sign for the mean is chosen randomly. Outliers were generated in two different ways.

Observation outliers: For an outlying observation, all variables of this observation are simulated as outliers, as described above. As a consequence, the outlier is indeed very different from regular observations, and it will thus have a severe impact on non-robust estimation.

Cell outliers: Outliers are generated only in randomly selected cells of the data matrix. Thus, depending on the total amount of outlying cells, each observation may contain outlying cells. This situation is more difficult to deal with.

3.3.3 Performance evaluation

This simulation study contains d variables from which d_0 are true biomarkers and $d_1 = d - d_0$ non-biomarkers. The performance of the method is evaluated in the same sense as a multiple testing procedure (Dunnett, 1955). In this case, the null hypothesis would be that a certain variable is not a biomarker. Then the TP (True Positives) denote the number of correctly identified biomarkers, the TN (True Negatives) are the number of correctly identified non-biomarkers, FP (False Positives) are the number of non-biomarkers that were declared as biomarker, and FN (False Negatives) are the number of biomarkers that were not identified as such. These numbers are also presented in Table 3.2, and they are the basis for computing the performance measures considered here: the True Positive Rate (TPR) and the False Discovery Rate (FDR), defined as:

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \qquad \mathrm{FDR} = \mathrm{FP}/(\mathrm{TP} + \mathrm{FP}).$$

The TPR reflects the proportion of true biomarkers which were correctly identified. Ideally, this value should be one. The FDR mimics the concept of the type-I error in hypothesis testing. Out of all decisions for biomarkers ("discoveries"), it provides the proportion on all wrong decisions. Ideally, the value for FDR should be zero.

Table 3.2: Classification of an outcome of biomarker identification: number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

	Identified as		
	Biomarker	Non-biomarker	Sum
True biomarker	TP	FN	d_0
Non-biomarker	FP	TN	d_1

3.3.4 Methods for comparison

The simulation design for the uncontaminated data follows closely the paper Filzmoser and Walczak (2014), and there it turned out that probabilistic quotient normalization (PQN) (Dieterle et al., 2006) had the best performance, followed by the method PLR which uses all pairwise log-ratios of the variables. All other investigated methods had some difficulties. For this reason, the PQN method will be used below as a reference method.

After PQN is carried out, a decision on significance of the variable, referring to a biomarker, has to be made. While in our procedure this decision is based on a critical

value (V_j^* from Equation (3.5) exceeds the threshold of the standard normal quantile $u_{0.975}$), in Filzmoser and Walczak (2014) this decision was based on the uninformative variables elimination partial least-squares method (UVE-PLS) (Centner et al., 1996; Zerzucha and Walczak, 2012). Basically, UVE-PLS borrows the ideas of a permutation test, since the observations of the original $\mathbf{X}$ matrix are randomly permuted. This is done several times, and all resulting matrices, multiplied by a small factor to reduce the importance, are augmented to $\mathbf{X}$, yielding a wide input matrix for PLS, which can be dealt with using the kernel PLS algorithm. The response for PLS is the class membership variable, and the stability of the regression coefficients is determined by leave-one-out cross-validation (Centner et al., 1996; Zerzucha and Walczak, 2012).

Note that in our studies where the data are contaminated by outliers, UVE-PLS may lead to biased results because of the non-robustness of PLS to outliers. This could be solved by using a robust partial least-squares method, like partial robust M regression (PRM) (Serneels et al., 2006). However, because of the high dimensionality of the augmented input matrix, and the computational complexity of the strategy to obtain significance, this would result in a very time consuming procedure.

On a standard PC, UVE-PRM would take around 400 minutes for one dataset we used in our simulations. Since for each parameter setting and each percentage of outliers in the data, 100 simulations were performed, and we also consider different numbers of observations in the groups, denoted as balanced and unbalanced settings, there are in total 22 400 simulations. Therefore, the computational time for all the simulations using UVE-PRM would be enormous. Thus, we will use UVE-PLS in combination with PQN in our comparisons.

In addition, we compare with two further well-known methods: DESeq2 (Love et al., 2014) and ALDEx (Fernandes et al., 2013; Gloor and Reid, 2016). Both are using an internal normalization of the data and its own biomarker identification decision. DESeq2 performs an internal normalization based on geometric means of each variable across all samples. Each sample is then divided by its mean. The median of the ratios is an estimation of the size-effect for a specific sample. Then a negative binomial generalized linear model is fitted for each variable, and the p-value from a Wald test (Wald, 1943; Harrell, 2014) is used for biomarker identification. This method also uses Cook's distance to detect outliers and removes them from the analysis. The method ALDEx generates Monte Carlo samples of the Dirichlet-multinomial model to derive the size-effect from the data. Then, the method internally applies the centered logratio transformation (Aitchison, 1986) and computes Welch's t-test for the biomarker identification. The authors of both methods recommended to adjust the p-values for multiple testing using the procedure of

Benjamini and Hochberg, see (Benjamini and Hochberg, 1995). In our simulations, an α level of 0.05 was used as a cut-off value.

3.3.5 Results for the balanced setting

In the balanced setting, both groups have the same number of observations, namely $n_1 = n_2 = 20$. As noted above, the number of variables is $d = 500$ and the number of true biomarkers is $d_0 = 20$, for each parameter setting listed in Table 3.1, 100 data sets were generated. In the following we present the averages for TPR and FDR over all simulations across all 8 parameter settings, and the corresponding standard errors are represented in the figures as gray areas.

Table 3.3 shows the results for the uncontaminated data. The first three methods are based on rPLR with the V_j^* statistic, see Equation (3.5), by using the empirical standard deviation (SD), the median absolute deviation (MAD), and the τ estimator of scale, respectively, to estimate the involved variation matrices. The last three methods of comparison are PQN in combination with UVE-PLS, as proposed in Filzmoser and Walczak (2014), DESeq2 (Love et al., 2014), and ALDEx (Fernandes et al., 2013). The performance of all methods is excellent, with slightly better results for the true positive rate for the last two methods. Note that other methods listed in Filzmoser and Walczak (2014) had difficulties with these situations.

Table 3.3: Average of the true positive rates (TPR), false discovery rates (FDR) and Standard Error of TPR over all simulations without outliers for the balanced setting, compared for the estimators SD, MAD, and τ for computing the V_j^* statistic, UVE-PLS combined with PQN, DESeq2 and ALDEx.

Method	TPR	FDR	SE of TPR
SD	0.986	0.000	0.0022
MAD	0.985	0.003	0.0024
τ estimator	0.988	0.000	0.0019
UVE-PLS (PQN)	0.985	0.006	0.0018
DESeq2	0.994	0.007	0.0053
ALDEx	0.994	0.009	0.0100

Figure 3.1 presents the results for observation outliers. Up to 50% of the observations from both groups are contaminated according to the scheme outlined in Section 3.3.2. Although such high outlier percentages are unrealistic in real data, it is interesting to see that the robust estimators MAD and τ still yield excellent results. The classical

standard deviation estimation SD gives reasonable results only for small contamination, but gets worse when the percentage of outlying observations increases. UVE-PLS for normalized data based on PQN is unreliable already for 10% outliers (4 observations). With increasing outlier percentage, also the results for DESeq2 and ALDEx deteriorate.

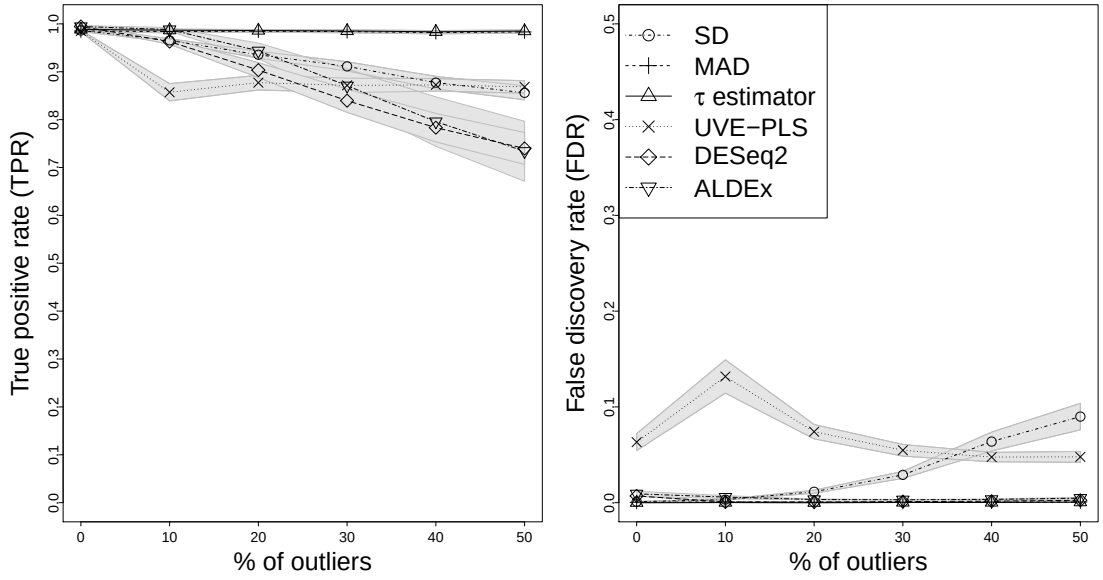


Figure 3.1: Observation outliers in the balanced setting: A given percentage (horizontal axes) of the observations is contaminated by outliers. The resulting TPR and FDR are averages over all simulation scenarios. The shaded areas represent the standard errors of the corresponding averages. Compared are different estimators of scale (SD, MAD, τ) for the V_j^* statistic, UVE-PLS based on PQN, DESeq and ALDEx.

Figure 3.2 shows the averaged TPR and FDR over all simulation scenarios for the balanced setting for cell outliers (see Section 3.3.2). The percentage varies from zero to 25%, the latter corresponds to 125 contaminated cells on average in each row of a data matrix, which is again quite extreme. Still, rPLR with the τ estimator as robust measure of scale delivers excellent results in this situation. The MAD leads to poor results of the TPR after including more than 12.5% outlying cells. The non-robust SD is not suitable in presence of outlying cells, and UVE-PLS with PQN leads to poor results for the TPR, but also for FDR in case of higher percentages. The true positive rates of DESeq and ALDEx are strongly influenced even by small percentages of cell outliers, while keeping the false discovery rate relatively low.

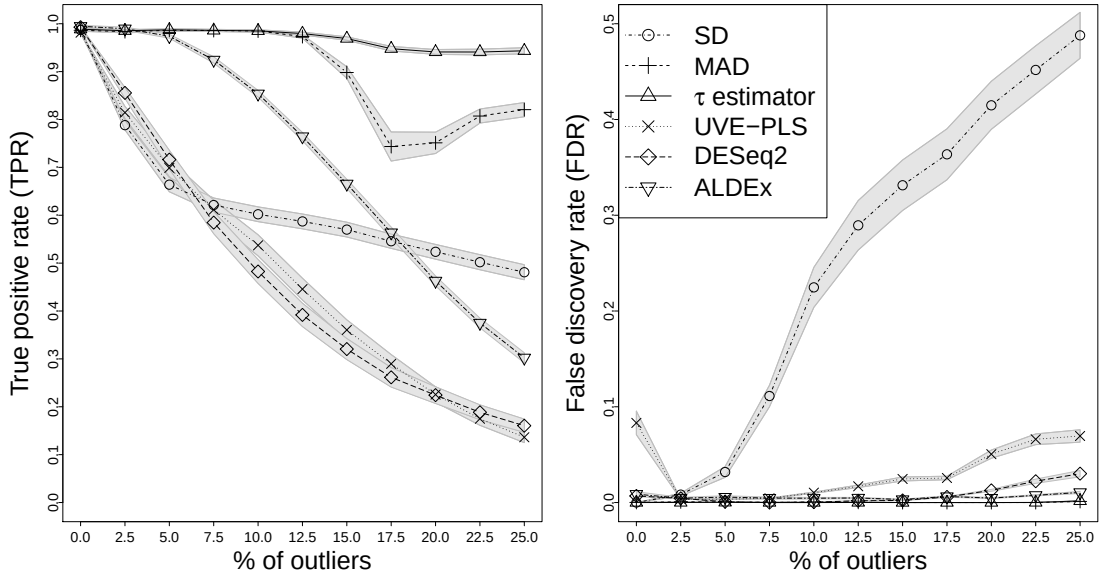


Figure 3.2: Cell outliers in the balanced setting: A given percentage (horizontal axes) of randomly selected cells of the data matrix is contaminated by outliers. The resulting TPR and FDR are averages over all simulation scenarios. The shaded areas represent the standard error of the corresponding averages. Compared are different estimators of scale (SD, MAD, τ) for the V_j^* statistic, UVE-PLS based on PQN, DESeq and ALDEx.

3.3.6 Results for the unbalanced setting

The number of observations in the groups is now fixed with $n_1 = 40$ and $n_2 = 5$, corresponding to a highly unbalanced situation. The number of variables and of true biomarkers is the same as before ($d = 500$, $d_0 = 20$). We consider the (more difficult) situation of cell outliers, where the outlying cells are now placed only in the observations of the bigger group. Note that we could also contaminate the smaller group, but since the cells are selected randomly, high percentages of outlying cells could completely destroy the information contained in the small group.

Figure 3.3 shows the results, again as averages of all 100 simulations for all 8 parameter combinations listed in Table 3.1. Already in the uncontaminated case (0% outlying cells) one can see important differences: UVE-PLS for PQN leads to a considerably lower TPR. The reason is that PLS is not appropriate for unbalanced groups. Also the MAD leads to a poor performance, in particular for the FDR. This is caused by the low efficiency of this estimator, which becomes crucial in the case of very small sample sizes. As expected, SD and UVE-PLS with PQN get worse in case of contamination. The proposed method with the τ estimator of scale gives excellent results for up to 17.5% ; then its performance

gets worse. Similar as in the balanced situation with cell outliers, DESeq2 and ALDEx have low TPR. On the other hand, these methods behave well in the situation without outliers in the unbalanced case.

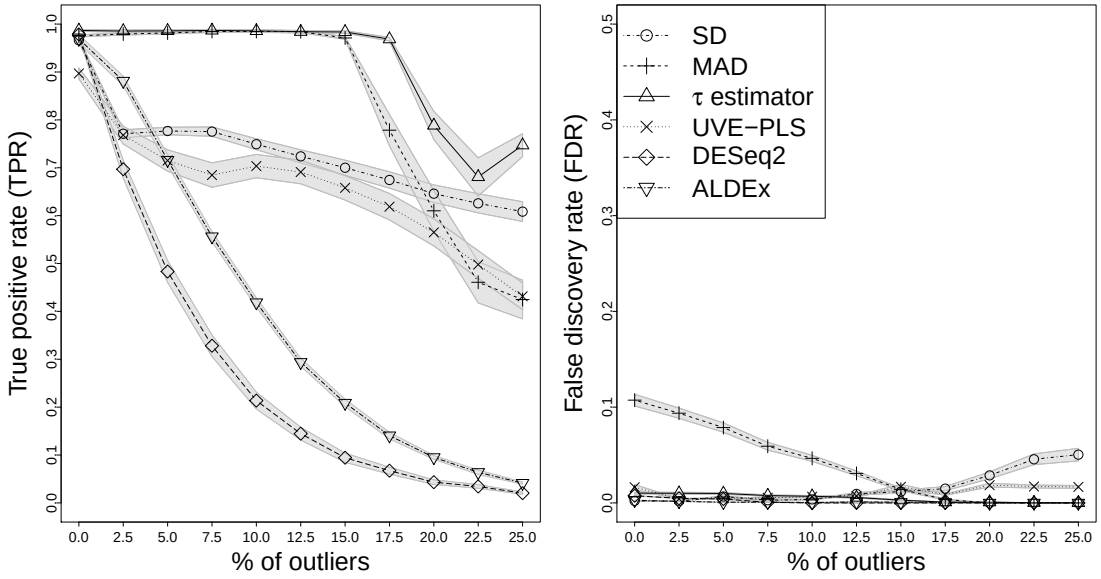


Figure 3.3: Cell outliers in the unbalanced setting: A given percentage (horizontal axes) of randomly selected cells of the observation in the larger group is contaminated by outliers. The resulting TPR and FDR are averages over all simulation scenarios. The shades areas represent the standard error of the corresponding averages. Compared are different estimators of scale (SD, MAD, τ) for the V_j^* statistic, UVE-PLS based on PQN, DESeq and ALDEx.

3.4 Outlier diagnostics

Identifying the correct biomarkers even in presence of moderate data quality is very desirable. However, it is also interesting to get more insight into the structure of potential data problems. In particular, it is interesting to know which observations are deviating from the majority in order to identify possible measurement errors or other artifacts. Also, identifying some variables or variable groups that show different behavior in all or parts of the observations may lead to important insights. For all these reasons, an outlier diagnostics tool is presented and discussed in this section.

The outlier diagnostics is based on the group variation matrices using the τ estimator of scale. The τ estimator internally computes weights w_i for the observations, see Equation (3.7). In our context, the input for the τ estimator are log-ratios of a pair of

variables,

$$\ln\left(x_{1j}^{(l)}/x_{1k}^{(l)}\right), \ldots, \ln\left(x_{n_l j}^{(l)}/x_{n_l k}^{(l)}\right) \tag{3.10}$$

see Equation (3.2), and the resulting estimated variance is $\hat{t}_{lk}^{(l)}$, the corresponding element of the variation matrix for the lth group ($l = 1, 2$). The weights of the τ estimator are thus assigned to each term in (3.10), leading to n_l weights. Since all variable pairs $j, k = 1, \ldots, d$ are considered for estimating the variation matrix, one can store all weights in a three-way array $W^{(l)}$ with d rows, d columns, and n_l slices. Denote the elements of this matrix by $w_{jki}^{(l)}$. Note that these robustness weights are computed already as part of the computation of the τ estimator, and thus the only additional effort is to store the weights.

The weights are used to identify cell-wise outliers, i.e. single matrix elements of the data matrix. Therefore, the information contained in $W^{(l)}$ needs to be aggregated appropriately. The information for a specific observation is contained in one particular slice of $W^{(l)}$, and due to the construction, this slice of dimension $d \times d$ is symmetric. We propose to average all weights for each observation and each involved variable,

$$m_{ij}^{(l)} = \frac{1}{d} \sum_{k=1}^{d} w_{jki}^{(l)}, \tag{3.11}$$

for $j = 1, \ldots, d$, $i = 1, \ldots, n_l$, and $l = 1, 2$. This information is stored in the $n_l \times d$ matrix $\mathbf{M}^{(l)}$, which can be represented graphically. All values are in the interval $[0, 1]$, where small values indicate outlying cells.

The outlier diagnostics is presented here for one simulated data set according to the simulation design of Section 3.3. We use the balanced situation with $n_1 = n_2 = 20$, and the parameter setting 4 from Table 3.1. The first 20 variables are simulated as biomarkers. Outliers are included cell-wise according to the black spots and fields in Figure 3.4 (upper plot), which represents the structure of the simulated data matrix. Thus, the cell-wise outliers are not only arranged randomly, but also in specific rows and parts of the matrix to test the method for diagnostics.

The lower part of Figure 3.4 shows the information of $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$, arranged on top of each other to obtain the same matrix dimension as for the upper plot. The weights are represented by a continuous gray scale, where a weight of zero corresponds to black, and a weight of one to white. One can see that all cells of the outlying rows were correctly identified, and also most cells of the outlying block. Also, most cell-wise outliers are correctly identified. Some cells are incorrectly indicated as outliers. However, due to the data generation, it is likely that some cells, although generated by normal distribution, are extreme just by chance.

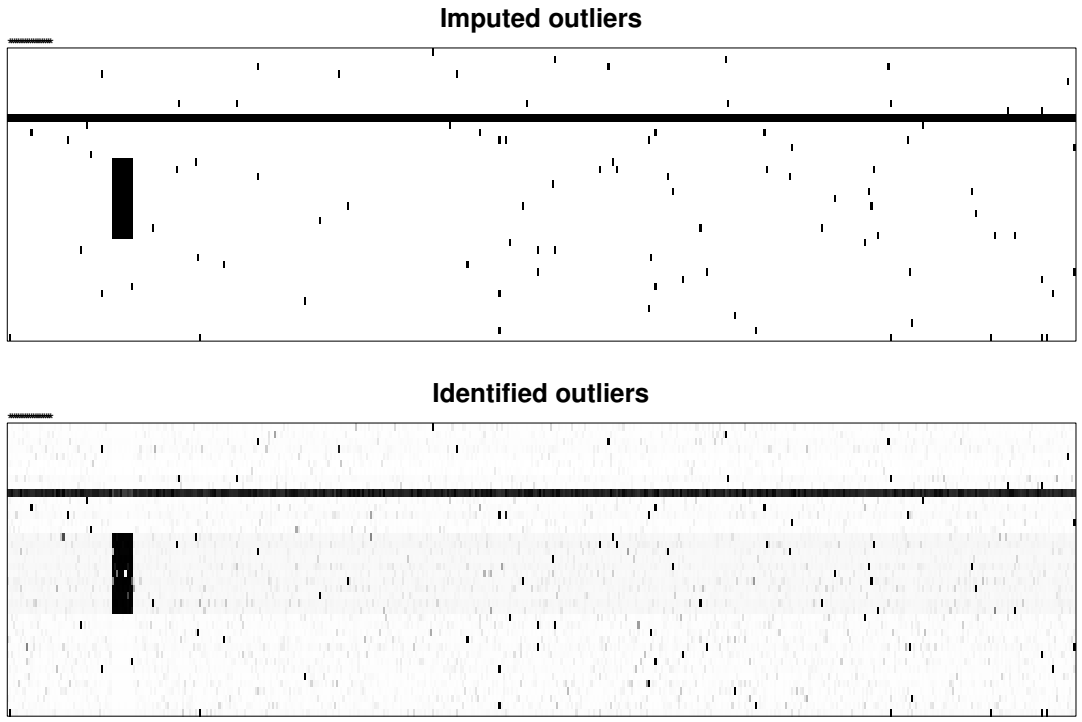


Figure 3.4: Outlier diagnostics for a simulated data set: true outlying cells in black (upper plot), and identified outlying cells (lower plot). The * symbols on the top of the plots represent the true biomarkers.

3.5 Example

Medium chain acyl-CoA dehydrogenase deficiency (MCADD OMIM # 201450) is one of the metabolomic diseases called the fatty acid oxidation disorders (FAODs). Blood samples from 25 healthy newborns as a control group and 8 newborns with MCADD disease were collected in the laboratory for inherited metabolic disorders (University Hospital Olomouc, CZ) within the pilot project of the Czech newborn screening program. Using subsequent sampling, the number of samples of patients suffering from MCADD was increased to 25. All these samples were analyzed using untargeted MS-based metabolomics.

The data were preprocessed by Laboratory of metabolomics (Institute of Molecular and Translational Medicine Faculty of Medicine and Dentistry Palacky University, Czech Republic). The data preprocessing was done in the software environment R (R Core Team, 2018) with the packages $XCMS$ (Smith et al., 2006) and $CAMERA$ (Kuhl et al., 2012). The $XCMS$ package was used for peak finding and 1900 features were identified.

In the next step, isotopes and adducts were grouped by the *CAMERA* package and excluded from the dataset. Quality control-based robust LOESS (LOcal regrESSion) signal correction was applied. During the preprocessing, the data dimension was reduced to 273 features. More details about the preprocessing can be found in Najdekr et al. (2015).

The resulting V_j^* values of rPLR using the τ estimator of scale of the described data are shown in Figure 3.5. By previous studies of MCADD, three biomarkers are known (Najdekr et al., 2015) (plotted as triangles). All the values above the plotted cut-off are identified as biomarkers by the new method. The identified biomarkers by UVE-PLS (PQN) are plotted as full circles. Altogether, UVE-PLS (PQN) identified 75 biomarkers out of 273 variables, and similarly, DESeq2 identified 99 and ALDEx even 155 biomarkers. These high numbers are rather unrealistic, since the biomarkers should represent only a small fraction of metabolites. These increased numbers corresponds to the simulation study with outliers, where the number of false positives was also quite high. On the other hand, the new method found all three previously known biomarkers as well as 11 additional metabolites. These biomarkers need to be investigated in detail by experts.

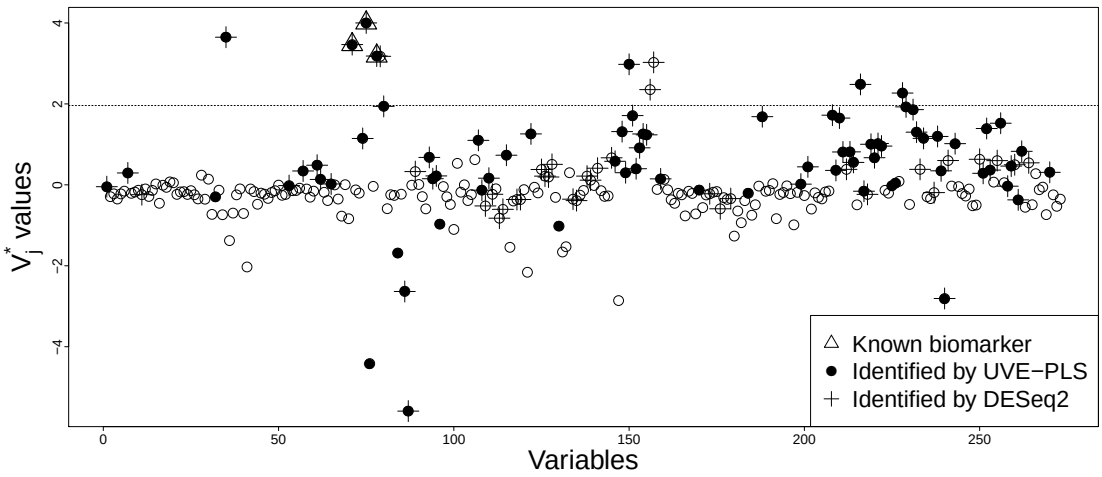


Figure 3.5: Biomarker identification: The new method identifies V_j^* values bigger than plotted cut-off as biomarkers. Full dots represent variables identified by UVE-PLS (with PQN) and plus signs represent variables identified by DESeq2. The results of ALDEx are not included in the plot since almost every variable (155 out of 273) was identified as biomarker. The triangles are known biomarkers by other studies of the disease.

To examine the data more deeply and to demonstrate the fact that a robust method should be used, the outlier diagnostics as described in Section 3.4 is performed, see

Figure 3.6. Several observations can be seen as outliers, especially observation number 3, 10, and possibly even 26, 27. On top of that, various cell outliers are present in the dataset. This suggests that the classical non-robust method might be influenced by the outliers, which might have spoiled the results of biomarker identification.

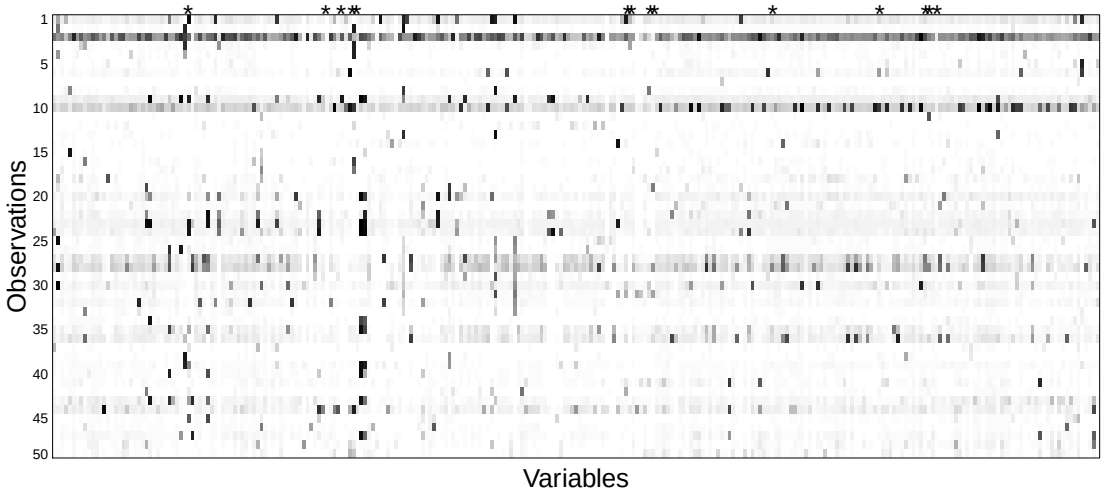


Figure 3.6: Outlier diagnostics in the real dataset MCADD. Observations 1 to 25 correspond with Control group, 26 to 50 with MCADD patients. The darker the cells, the higher is the probability of outlyingness. Identified biomarkers are indicated by *.

3.6 Conclusions

In this paper, a novel robust method (rPLR) is proposed for the identification of the key features (biomarkers), which are able to distinguish between two groups of samples (e.g. patients with and without certain condition). This method can handle data where the so called size effect is present. There are several methods suited to deal with this problem, however, to the best of our knowledge, there is no robust method available which is capable of dealing with outliers.

The new method is based on the variance of pairwise log-ratios. These log-ratios are scale invariant, so they are suitable for dealing with the size-effect problem. The method can be easily robustified just by estimating the variance (or its square-root) robustly. Here, three versions of the method are presented, which are based on the following scale estimators: 1) the non-robust standard deviation, 2) the robust median absolute deviation, and 3) the robust τ estimator of scale.

Simulation studies were carried out to investigate the performance of the new method under different scenarios and to compare with UVE-PLS with the PQN transformation, DESeq2 and ALDEx. The focus of the comparison was especially on the robustness behavior. The simulations were evaluated based on the true positive rate (TPR) and the false discovery rate (FDR). Two settings, the balanced and the unbalanced case were considered. Also, two different types of outliers were generated: observational outliers and cell outliers. Overall, UVE-PLS with the PQN transformation leads to good results if there are no outliers and if the groups are balanced, but the results get worse with increasing contamination and if the group sizes are very different. This is not directly related to the PQN transformation, but rather to UVE-PLS which is sensitive to outliers and unbalanced groups. The methods DESeq2 and ALDEx behave well in unbalanced situations, but they fail if outliers are present in the data. The simulation also verified that cell-wise outliers are more difficult to deal with compared to observational outliers. The new method based on log-ratios works well for the unbalanced case. In presence of outliers, a robust scale estimator is indispensable. The best results were achieved for the τ scale estimator due to its high breakdown point and high efficiency. In this paper, we evaluated the methods based only on TPR and FDR using certain cutoff values. We did not consider the frequently used Area Under the Receiver Operating curve (AUC). In our context, when only a small number of biomarkers is present, the AUC will depend mainly on the evaluation of non-biomarkers. For this reason AUC results are not included here. Anyway, AUC would suggest similar conclusions as using TPR and FDR. The τ scale estimator has another advantage: The weights which are internally computed to downweight outlying observations can be aggregated to identify outlying cells in the data matrix. This information has been exploited in a diagnostics plot. Such a plot can provide valuable information about data artifacts, since it reveals if complete observations have deviating data structure, or if certain variables or cells have different behavior. The reasons for outlyingness can be manifold, including problems during the measurement or preprocessing step.

We believe that the plot of the V_j^* values, see e.g. Figure 3.5, and the diagnostics plot, see Figure 3.6, provide valuable insight into the data analysis. The analyst may not only be interested if a variable is a biomarker or not, but it is also interesting to see, to which "degree" a variable is identified as biomarker. This is visible in the plot of the V_j^* values. The new method is implemented in the R package `robCompositions` as function `biomarker`, yielding both plots as an output.

It should be admitted that the proposed method has limitations concerning zero entries in the data matrix. Zeros in combination with log-ratio methods lead to values of

infinity, and thus to numerical difficulties. A way out would be to impute the zero values, which might be considered as values below a certain detection limit, see Templ et al. (2016), but currently it is not clear how the performance of the method would be affected, and thus further research is needed in this direction. Another limitation is the fact that the method would be computationally demanding in extremely high dimensions. In our future research we also plan to extend this work from the two-group to the multi-group setting.

Acknowledgments

This research is supported by the Austrian Science Fund (FWF) and Czech Science Fund (GACR), project number I 1910-N26.

CHAPTER 4

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

Abstract: Data outliers can carry very valuable information and are often most informative for the interpretation. An algorithm called cell-rPLR for the identification of outliers in single cells 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 weights results in outlyingness information. A further feature of cell-rPLR is that it is useful for biomarker identification, particularly in presence of cellwise outliers. Real data examples and simulation studies underline the good performance of this algorithm in comparison to alternative methods.

4.1 Introduction

Metabolomic data, as well as many other data sets from "omics" disciplines, are high-dimensional, with many variables and commonly limited by few observation, originating

from two or more different groups (controls, 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 (Strimbu and Tavel, 2010; Pepe et al., 2001).

Biomarker identification is getting more challenging if outliers are present in the data (Abeel et al., 2009). Outliers in this context can be observations that are somewhat different in their data structure compared to the data majority, and this difference may be caused by measurement problems, different data preprocessing, inconsistencies among the observations, etc (Huber and Ronchetti, 1981; Maronna et al., 2019). 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, i.e. 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 4.1 (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, see Maronna et al. (2006). In contrast to that, Figure 4.1 (right) refers to a scheme of "cellwise" outliers, where single cells of the data matrix (colored in black) are identified as outliers. 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 (Rousseeuw and Bossche, 2018), as well as the development of robust estimators with cellwise outliers (Öllerer et al., 2016).

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 with different sample volume. Thus, the obtained data values are not directly comparable, and the data need to be preprocessed first before applying a statistical method (Warrack et al., 2009; Filzmoser and Walczak, 2014). Preprocessing can be done by making use

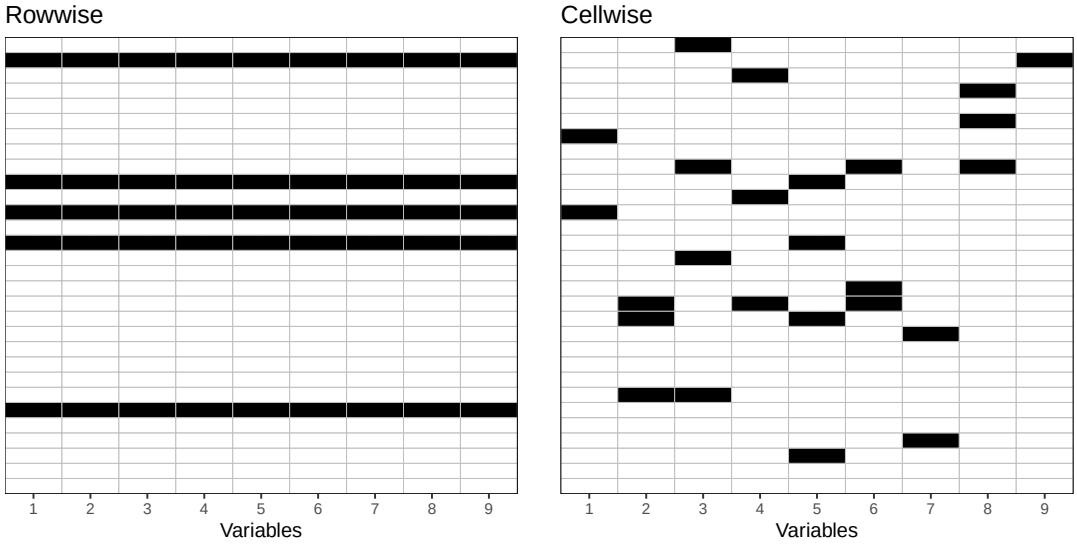


Figure 4.1: Difference between rowwise (left) and cellwise (right) outliers of a data matrix.

of a specific data transformation or normalization, e.g. the total-sum normalization (TSN) (Craig et al., 2006) or the probabilistic quotient normalization (PQN) (Dieterle et al., 2006). An alternative is to use the log-ratio methodology from compositional data analysis, which is based on pairwise log-ratios (Pawlowsky-Glahn et al., 2015). Since for two observations $\mathbf{x}$ and $\mathbf{y}$, and any positive constant s (representing the size effect), $\ln\left((s \cdot \mathbf{x}) \,/\, (s \cdot \mathbf{y})\right) = \ln\left(\mathbf{x}/\mathbf{y}\right)$, preprocessing is not necessary here (Walach et al., 2018).

In this paper we will introduce an algorithm called *cell-rPLR*, which is the abbreviation for cellwise outlier diagnostics using robust pairwise log-ratios. The goal of cell-rPLR is twofold: it can be used for (a) cell-wise outlier identification and for (b) biomarker identification. Section 4.2 describes the theoretical basis of the method. Section 4.3 introduces a diagnostics plot for cellwise outlier identification. In a simulation scenario, cell-rPLR is compared to an alternative approach for cellwise outlier detection. Section 4.4 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 4.5 summarizes the paper, provides information about software, and concludes.

4.2 Method

Let us assume a dataset arranged in a data matrix $\mathbf{X}$, with n samples and d variables. The matrix $\mathbf{X}$ 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 $n_1 + \ldots + n_g = n$.

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 a weighting 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.

4.2.1 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:

$$\ln\left(\frac{x_{1j}^{(1)}}{x_{1k}^{(1)}}\right), \ldots, \ln\left(\frac{x_{n_1 j}^{(1)}}{x_{n_1 k}^{(1)}}\right), \ln\left(\frac{x_{n_1+1,j}^{(2)}}{x_{n_1+1,k}^{(2)}}\right), \ldots, \ln\left(\frac{x_{n_1+n_2,j}^{(2)}}{x_{n_1+n_2,k}^{(2)}}\right), \ldots, \ln\left(\frac{x_{nj}^{(G)}}{x_{nk}^{(G)}}\right) \tag{4.1}$$

Clearly, the log-ratios are zero if $j = k$, and exchanging denominator and nominator leads to the same log-ratio, but with different sign. Subsequently, we will assign a weight to each entry of the pairwise log-ratios. In order to design an appropriate weight function, the log-ratios need to be centered and scaled first. Since potential group-differences should not get lost, centering and scaling is performed with respect to the entries in one group only, where we propose to use the majority group for this purpose. 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 $1 < g \leq G$. Further, we simplify the notation by defining $y_{ijk} := \ln\left(\frac{x_{ij}^{(g)}}{x_{ik}^{(g)}}\right)$, 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 (Maronna et al., 2006), as

$$\bar{y}_1 = \frac{\sum_{i=1}^{n_1} v_i y_i}{\sum_{i=1}^{n_1} v_i}, \tag{4.2}$$

where

$$v_i = \omega_c\left(\frac{y_i - \text{median}\,(y_1, \ldots, y_{n_1})}{s_1}\right), \tag{4.3}$$

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

$$\text{MAD}(y_1, \ldots, y_{n_1}) = 1.483 \cdot \text{median}_i(\mid y_i - \text{median}_j(y_j) \mid) \tag{4.4}$$

The function $\omega_c(\cdot)$ in Equation (4.3) is Tukey's biweight function (Beaton and Tukey, 1974), defined as

$$\omega_c(u) = \left(1 - \left(\frac{u}{c}\right)^2\right)^2 \cdot I\left(u, c\right), \tag{4.5}$$

with

$$I\left(u, c\right) = \begin{cases} 1, & \text{for } |u| < c. \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

The tuning constant is usually chosen as $c = 4.685$. For more details, we refer to Yohai and Zamar (1988) and Maronna and Zamar (2002), who introduced these concepts in the framework of robust scale estimation.

The robustly centered and scaled values are obtained as

$$\tilde{y}_i = \frac{y_i - \bar{y}_1}{s_1} \quad \text{for } i = 1, \ldots, n. \tag{4.7}$$

Centering and scaling is 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 (4.3) are not defined, because s_1 would be zero. 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.

4.2.2 Weighting functions

The robustly centered and scaled pairwise log-ratios contain information about outlyingness, and this information will be revealed by applying an appropriate weight function to these values. A weight function as proposed in Equation (4.5) would, however, not be appropriate, since the resulting weights 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

$$\omega_c^*(u) = \omega_c(u) \cdot sgn(-u) + sgn(u), \tag{4.8}$$

with the sign function

$$sgn(v) = \begin{cases} 1, & \text{for } v \geq 0, \\ \text{-1}, & \text{otherwise,} \end{cases} \tag{4.9}$$

yielding values in $[-1, 1]$. Figure 4.2 shows the original definition of the Tukey biweight weights (left plot) and compared with the adjusted version (right plot).

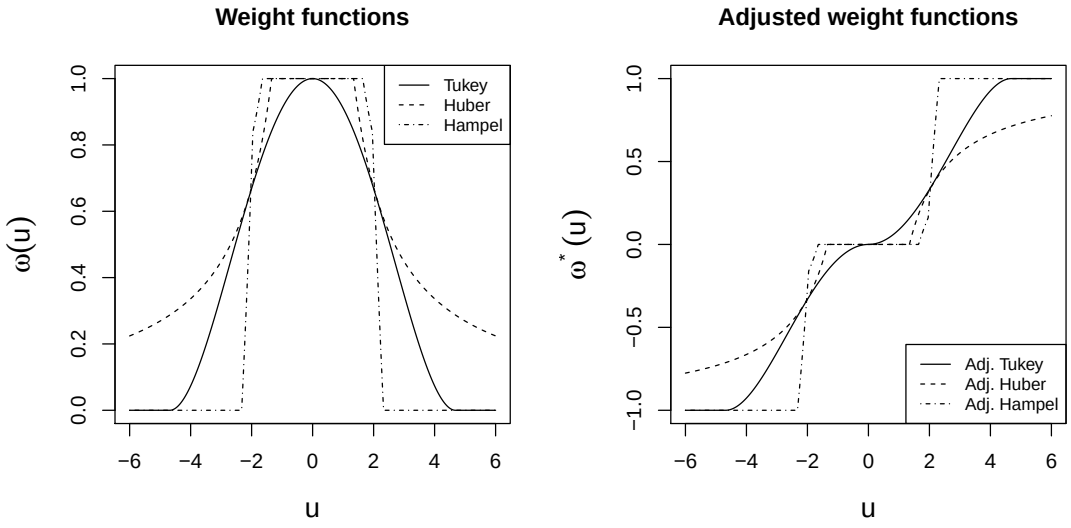


Figure 4.2: Original (left) and adjusted (right) weighting functions

The adjusted weight function is applied to the robustly centered and scaled pairwise log-ratios from (4.7). Weights around zero represent non-outlying values, and weights closer to -1 or $+1$ represent potential outliers.

The shape of the weight function will determine the characteristics of the outlier detection method, and thus other weighting functions shall be considered as well. In the literature on robust statistics, many proposals are available, such as Huber's (Huber and Ronchetti, 1981) and Hampel's (Hampel et al., 1986) functions. Below we will propose the adjusted versions, resulting in weights in the interval $[-1, 1]$.

Huber weighting function: The original definition is

$$\omega_k(u) = min\left(1, \frac{k}{|u|}\right), \tag{4.10}$$

where k is a tuning parameter, typically taken as 1.345 (Huber and Ronchetti, 1981). 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

$$\omega_k^*(u) = (\omega_k(u) - 1) \cdot sgn(-u) \tag{4.11}$$

assigns weights of 0 to those (non-outlying) log-ratios which are still in the usual range. The resulting shape of the adjusted Huber weighting function can be seen in Figure 4.2 (right); the left plot shows the original definition.

Hampel weighting function: The original definition is

$$\omega_h(u) = \begin{cases} 1 & |u| \leq c_1 \\ \frac{c_1}{|u|} & c_1 \leq |u| \leq c_2 \\ \frac{c_3 - |u|}{c_3 - c_2} \frac{c_1}{|u|} & c_2 \leq |u| \leq c_3 \\ 0 & c_3 < |u| \end{cases} \tag{4.12}$$

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 zero asymptotically, whereas in the Hampel function one obtains zero weights according to the tuning parameter c_3. The adjusted version of the Hampel function is

$$\omega_h^*(u) = (\omega_h(u) - 1) \cdot sgn(-u), \tag{4.13}$$

which again provides values in $[-1, 1]$. Figure 4.2 shows the original (left) and adjusted (right) Hampel weighting function.

4.2.3 Aggregation of weights

Let us denote the adjusted weight function by $\omega^*(\cdot)$, which refers to one of the proposed functions in Section 4.2.2. We apply this function to the centered and scaled pairwise log-ratios, see Section 4.2.1, resulting in weights

$$w_{ijk}^* = \omega^*(\tilde{y}_{ijk}) \tag{4.14}$$

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

Since we aim at a method for cell-wise outlier detection, the weights in the array $\mathbf{W}^*$ need to be aggregated appropriately in order to obtain weights for each cell in the $n \times d$ data matrix $\mathbf{X}$. For robustness reasons we propose the aggregation into weights

$$w_{ij} = \text{median}\left(w_{ij1}^*, w_{ij2}^*, \ldots, w_{ijd}^*\right), \tag{4.15}$$

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

4.2.4 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 (4.1).

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 a weighting function to $\tilde{y}_{ijk}$, which yields weights w_{ijk}^*, see (4.14).

Step 4: Aggregate the weights according to (4.15) to obtain the final weights w_{ij}, arranged in the weight matrix $\mathbf{W}$.

Note that this outlier detection algorithm is supervised, because the group information of the observations is used in Step 2.

We do not specify an outlier cut-off value for identifying outlying cells. Rather, we visualize the information contained in $\mathbf{W}$, by using different colors for positive (red) and negative (blue) values, and different color intensity, with light color for weights close to zero, and intense color otherwise. Thus, cell-rPLR serves as a visual outlier diagnostics tool.

4.2.5 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 weights w_{ij} from Step 4 of the algorithm are still associated to the groups, and since we arranged the observations group-wise, we have weights $w_j^{(1)} = \{w_{1j}, \ldots, w_{n_1 j}\}$ for the first group, and weights $w_j^{(2)} = \{w_{n_1+1,j}, \ldots, w_{nj}\}$ for the second group, for $j = 1, \ldots, d$. For feature selection we compare the medians in both sets of weights by

$$m_j = \left| \text{ median}\left(w_j^{(1)}\right) - \text{median}\left(w_j^{(2)}\right) \right|. \tag{4.16}$$

The larger the difference is, the more important the variable is for the discrimination of the groups. Note that the size of m_j for different j can indeed be compared, since the weights are on the same scale. One can either sort the values m_j in descending order, and 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 (Fisher, 1935; Rubin, 1980) are wildly 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, i.e. $m_j = 0$, and thus the variable is not a biomarker.

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

Step 1: Use as an input the matrix $\mathbf{W}^*$ with the elements w^*_{ijk}, defined in (4.14).

Step 2: Randomly permute the values w^*_{ijk} according to the index j, resulting in values $w^*_{ijk}(b)$ for replication $b \in \{1, \ldots, B\}$

Step 3: Aggregate the values from the b-th permutation as in Eq. (4.15), yielding values $w_{ij}(b)$.

Step 4: Compute the differences according to Eq. (4.16), resulting in $m_j(b)$, for $j = 1, \ldots, d$.

Step 5: Compute the proportion

$$\frac{1}{B}\sum_{b=1}^{B}(m_j \leq m_j(b))$$

for $j = 1, \ldots, d$, which is interpreted as p-value for the j-th variable. Here, m_j refer to the values from (4.16) for the unpermuted data.

In our numerical experiments we used $B = 1000$. The computations are still feasible, because the input matrix $\mathbf{W}^*$ is fixed, and with the B permutations, the p-values for all variables are returned.

Note that in Step 2, the permutations do not necessarily have to be done just in the index j, but one could permute all the elements in the blocks of the array $\mathbf{W}^*$ which correspond to the two groups of observations. The results would essentially be the same, since the test statistic (4.16) is based on group-wise medians of median-aggregated values, see (4.15). Since this is numerically easier to do, we have implemented this option.

4.3 Performance of cell-rPLR for outlier identification

The cell-rPLR algorithm results in weights w_{ij}, arranged in the weight matrix $\mathbf{W}$, see Equation (4.15) which indicate 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 colour information. We will use red color for positive weights, blue color for negative values, and the magnitude of the weight values will determine the color saturation (values around zero in light colors).

4.3.1 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 (Janečková et al., 2012) 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, Palacký University Olomouc, Czech Republic. There are in total four different inherited metabolic disorders (IMD) – phenylketonuria (PKU), homocystinuria (HCYS), methylmalonic aciduria (MMA), propionic aciduria (PA), each with different number of samples varying from two to six. 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 was processed in MetIQ software (AbsoluteIDQ kit). In total 163 metabolites were quantified.

MTBL59: This data set is described in Franceschi et al. (2012); Wehrens et al. (2011) and can be downloaded from the MetaboLight web page `https://www.ebi.ac.uk/metabolights/MTBLS59`. It contains twenty apple samples which were analyzed by Liquid chromatography mass spectrometry (LC-MS). The first ten samples of apples were analyzed without any modification. The second ten 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 pre-processing was carried out as outlined in Wehrens et al. (2011). Only the first nine mintes of the chromatography were 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 Wang et al. (2014) was performed in the Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacký University Olomouc, Czech Republic. The

data were processed using the vendor software Compound Discoverer 3.0 (Thermo Fisher Scientific), exported to the R software to perform correlation analysis to merge redundant features (RT difference ≤ 0.02 min, r ≥ 0.95, clr transformation) and to perform other statistical evaluation (the data were corrected using QC samples and LOESS regression (Cleveland and Devlin, 1988), potential metabolites with CV higher than 30% were excluded from further data processing). There are in total 2336 features in the data set. Based on the biological analysis of the PKU disease, there are four known biomarkers (Miller et al., 2015; Jansen et al., 2015; Václavík et al., 2018).

4.3.2 Visualization of cellwise outliers

We use the data set *IMD* to visualize cellwise outliers. In this multi-group data set, centering and scaling was performed according to the majority group, which is the control group, see Equation (4.7). Then the cell-rPLR method was applied as described in Section 4.2. The resulting heatmap is shown in Figure 4.3 using the adjusted Tukey biweight function (4.8), and in Figure 4.4 which is based on the adjusted Hampel function (4.12). For reasons of space, we omitted the first 36 control patients from the visualization, and show only Controls 37-50, and the patients from the different disease groups, separated by black horizontal lines. At a first glance, Figure 4.4 seems to provide a much clearer picture concerning potential cellwise outliers. This is due to the fact that the adjusted Hampel function assigns zero 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 which is higher than expected. A blue value indicates a "negative outlier", with a value lower than expected. Figure 4.3 shows that control observation 48 has systematic bias from the others: the first block of variables for this sample are outlying in the positive direction whereas the last block is outlying in the negative direction. This is not so clearly visible in Figure 4.4. A possible explanation for the outlyingness might be that the sample preparation was done a bit different from the other samples, a different device setting was used during the analysis or control might have been biased by 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) means that the corresponding variables have increased values

(dominance) for these observations.

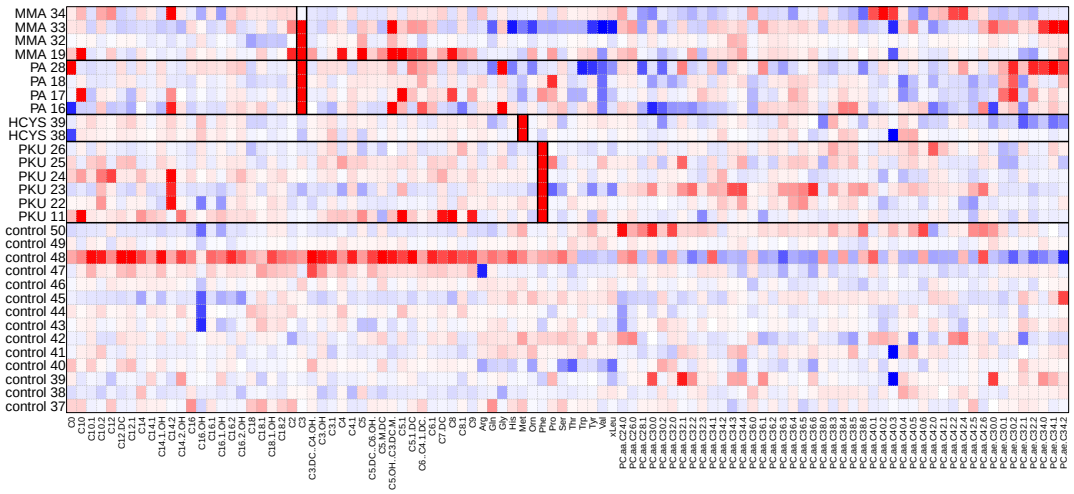


Figure 4.3: Outlier diagnostics for the *IMD* data, using the adjusted Tukey biweight function

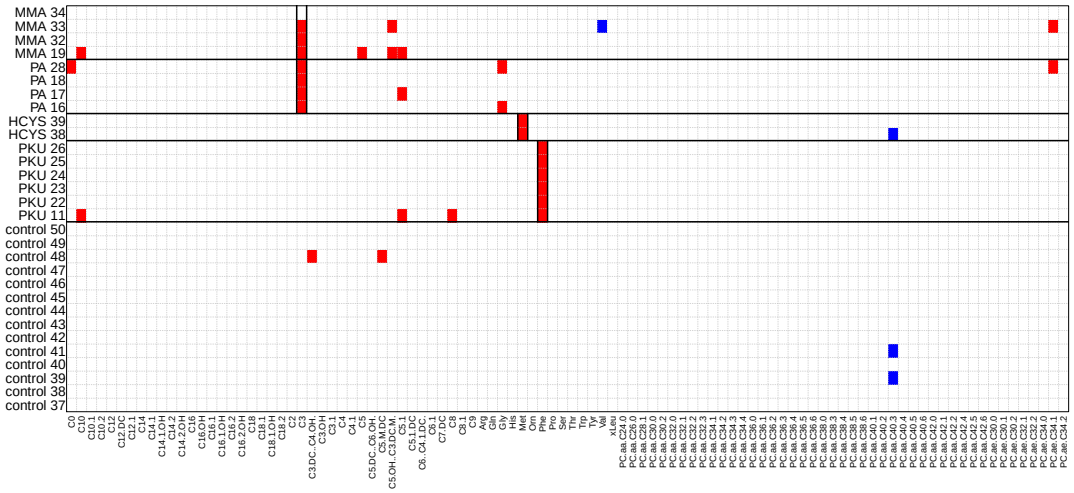


Figure 4.4: Outlier diagnostics for the *IMD* data, using the adjusted Hampel function

This example shows that the choice of the weighting 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 colors

instead of white cells , but indicate clearer small deviations from the "normal" behavior.

4.3.3 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

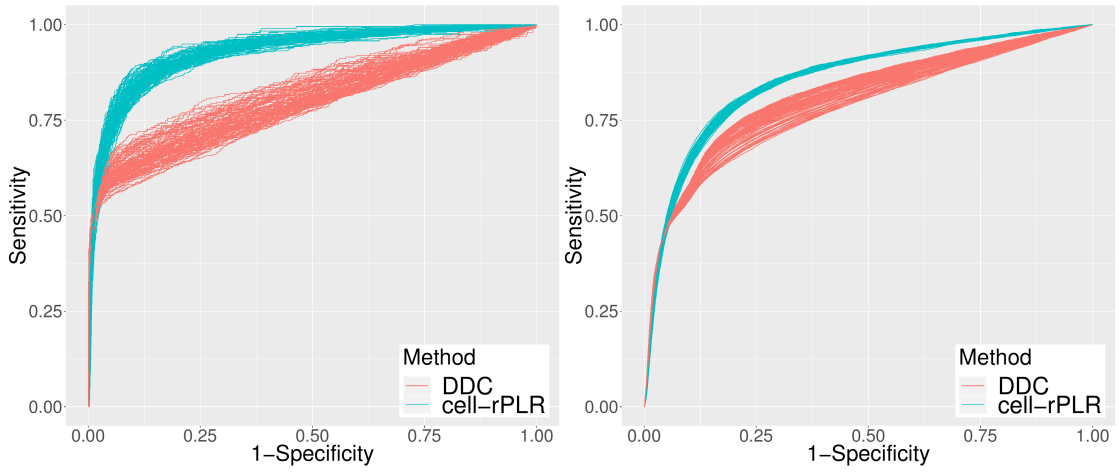
$$\tilde{x}_{ij} = x_{ij} \cdot M + A, \tag{4.17}$$

where a multiplicative effect M is generated from a uniform distribution $\mathcal{U}(0.2, 0.5)$ or $\mathcal{U}(2, 10)$ (with equal probability), and an additive effect A, generated from $\mathcal{U}(2, 5)$. 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 (4.17) could produce values which 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 DDC (Detect Deviating Cells),(Rousseeuw and Bossche, 2018), which can be considered as a state-of-the-art cellwise outlier identification method. In the first step, DDC robustly standardizes 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 non-marked observations. This is followed by the prediction of each non-marked 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.

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 non-outliers (one minus Specificity) for varying outlier cut-off points. Figure 4.5 shows the 100 different ROC curves for the cell-rPLR (with the adjusted Tukey biweight function) and the DDC algorithm for the data set (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



(a) *MTBL59* data (b) *PKU* data

Figure 4.5: ROC curve for the identification of cellwise outliers of the two algorithms DDC (red) and cell-rPLR (blue).

data sets, the cell-rPLR algorithm clearly shows a better performance compared to the DDC method.

4.4 Performance of cell-rPLR for biomarker identification

As in the previous Section 4.3.3, 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 presence of contamination by cellwise outliers. Therefore, the same kind of data contamination is applied as introduced in (4.17), 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 4.2.5. A ranked list for the variable importance is obtained by the comparison of the medians of the weights, see Equation (4.16). For a list of identified biomarkers, the permutation test as described in Section 4.2.5 is applied.

PLS-VIP: Partial least squares discriminant analysis (PLS-DA) (Wold, 1975; Wold et al., 1983; Ståhle and Wold, 1987) is applied to the data set, which is doing partial least squares (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 VIP (variable importance projection) score (Favilla et al., 2013; Wold et al., 1993) is computed which sums up the contribution of each variable in the model. The VIP scores are then sorted is 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 one (Chong and Jun, 2005; Gosselin et al., 2010).

PLS-SR: There are many methods available for extracting the importance of the variables in a PLS-DA model (Mehmood et al., 2012). Here we consider the Selectivity Ratio (SR) (Rajalahti et al., 2009a; Kvalheim, 2009) 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 $F_{n-2,n-3}$ (Rajalahti et al., 2009b).

PRM-VIP: PLS-DA is not robust against data outliers (Filzmoser et al., 2009), and thus the robust counterpart based on PRM (partial robust M-Regression) (Serneels et al., 2005) 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 Selectivity Ratio.

DesEq: This method builds on two steps: a) an internal normalization of the variables by their geometric means, b) a decision about the importance of variables (Love et al., 2014). For this purpose, a negative binomial generalized linear model is fit to each variable, and the p-value from a Wald test (Wald, 1943; Harrell, 2014) is computed for creating a rank of importance for the variables, and a list of identified biomarkers. The method removes outliers based on the Cook's distance.

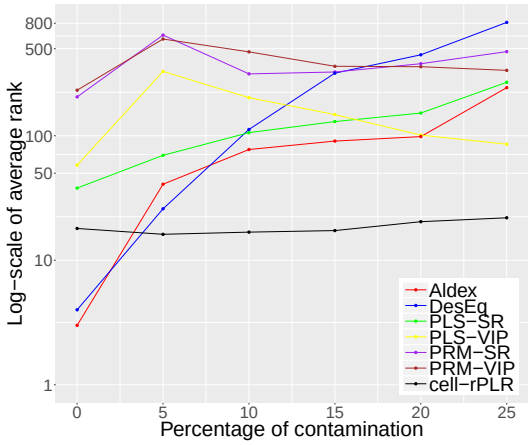
Aldex: This method is based on Monte Carlo simulations of the Dirichlet multinomial model (Fernandes et al., 2013; Gloor and Reid, 2016). The centered log-ratio transformation (clr) (Aitchison, 1982) is internally used. Then, p-values obtained from Welsch's test (Welch, 1938) are employed for ranking 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 (Dieterle et al., 2006) which is widely used.

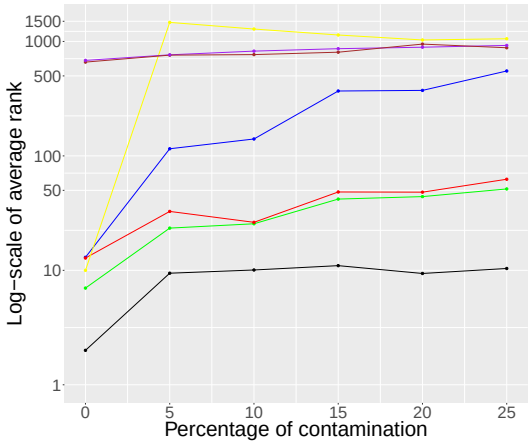
Figures 4.6 and 4.7 show the results of the simulation study. The subplots in these figures report the resulting ranks for one particular known biomarker. Figure 4.6 refers to the results for the *PKU* data with four known biomarkers, and Figure 4.7 to those from the *MBTL59* data with five known biomarkers. We report here the average ranks (over the 100 simulations) for the individual methods.

Figure 4.6 and 4.7 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 4.6), 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 4.7. 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 4.6), PLS-SR is also quite competitive, but it completely fails for the *PKU* data.

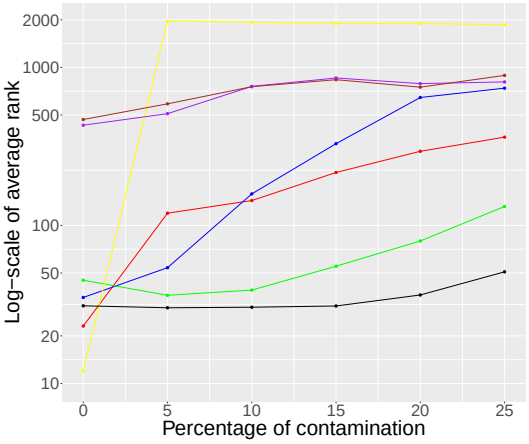
Each of the considered methods returns the information if a variable is identified as a biomarker or not. In case of cell-rPLR, the permutation test (see Section 4.2) is employed to deliver this information. Thus, we evaluate the performance of correct biomarker identification based on the same simulation 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 non-biomarkers (Specificity) is computed. In the ideal case, both the TPR and TNR should yield values close to one. Figure 4.8 summarizes the average values for TPR and TNR over the 100 simulation runs, and for the MTBL59 data set. According to the 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, as well as DexEq 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.



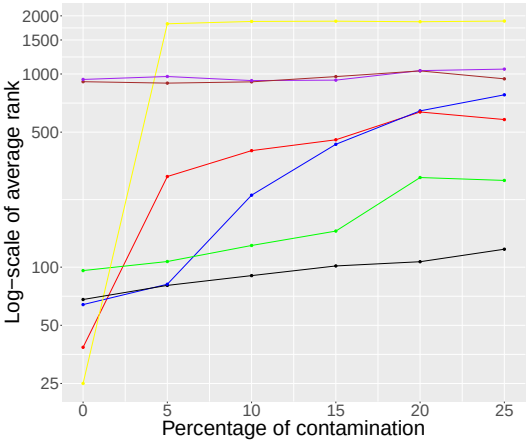
(a) Known biomarker #1



(b) Known biomarker #2

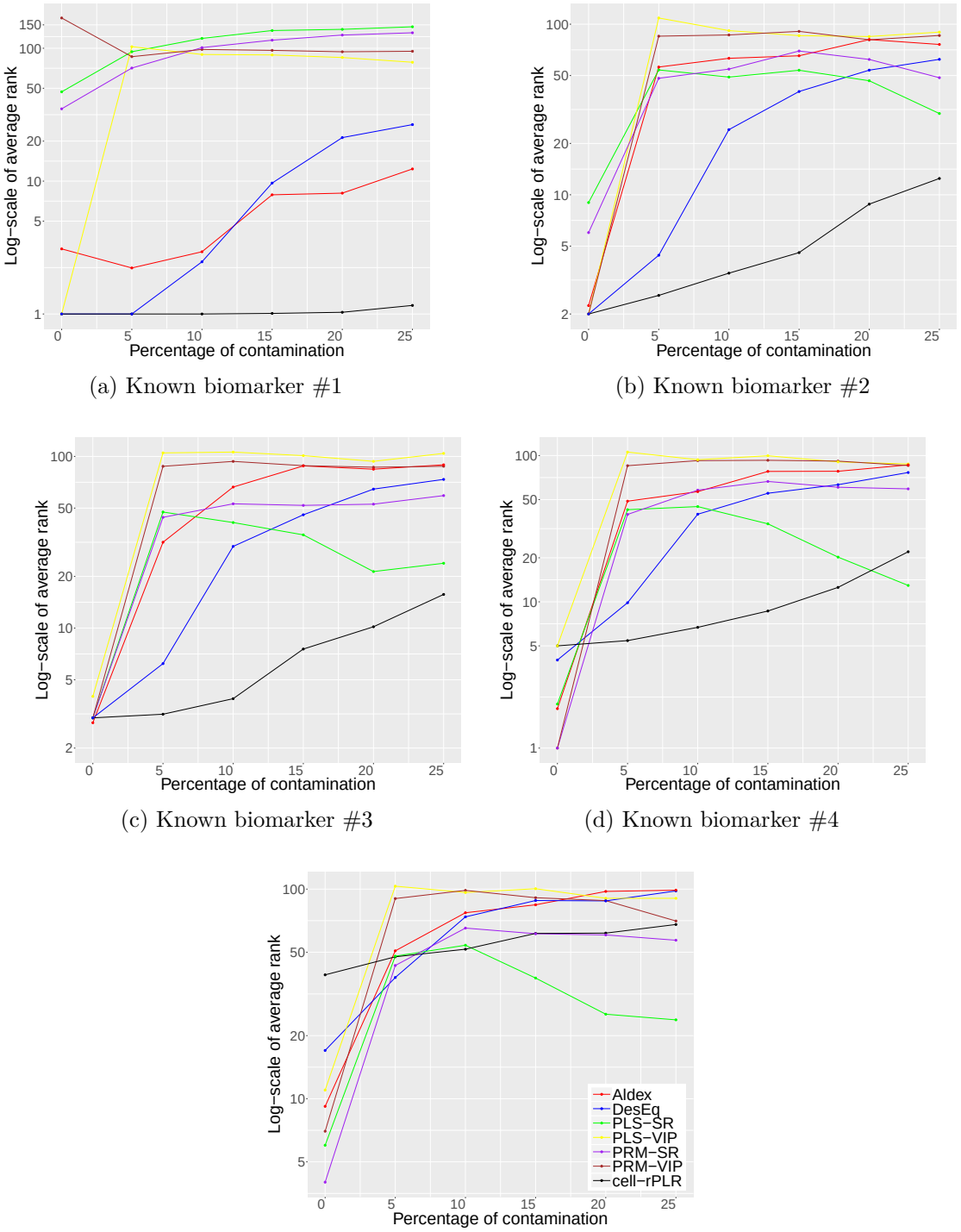


(c) Known biomarker #3



(d) Known biomarker #4

Figure 4.6: 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.



(a) Known biomarker #1

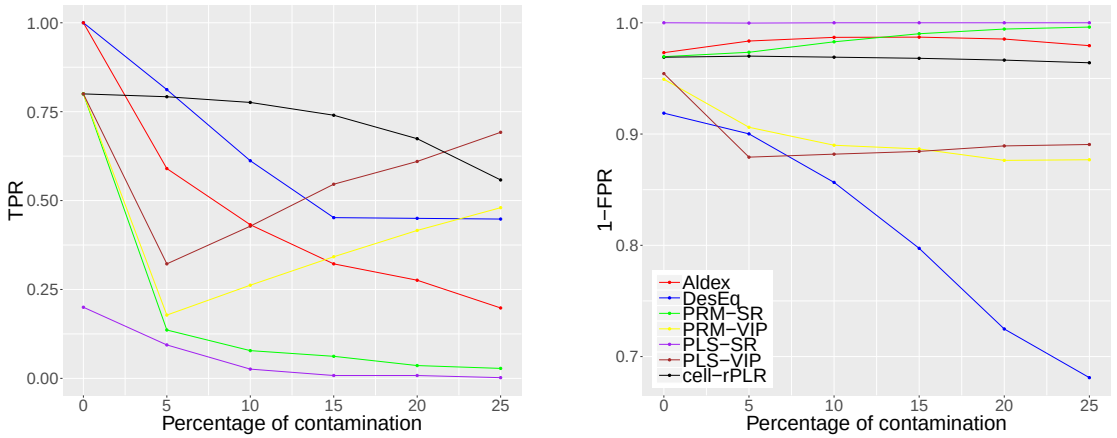
(b) Known biomarker #2

(c) Known biomarker #3

(d) Known biomarker #4

(e) Known biomarker #5

Figure 4.7: 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.



(a) True Positive Rate (TPR)

(b) True Negative Rate (TNR)

Figure 4.8: Performance of the methods for their ability in biomarker identification for the MTBL59 data set, for different levels of cellwise contamination.

4.5 Summary and conclusions

This paper introduced a method called cell-rPLR (cellwise outlier diagnostics using robust pairwise log-ratios), 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, scaling) needed by many other methods, because the elemental information are pairwise log-ratios between the variable values. The size effect which 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 a weight function to the robustly centered and scaled pairwise log-ratios, and the results are weights for each observation, and for each pair of variables (3-way array). After robust aggregation one obtains a weight matrix of the same dimension as the data matrix, containing the cell-wise outlyingness information. The weights are in the interval $[-1, 1]$; weights around zero point at "normal" data cells, and weights close to $+1$ or -1 indicate atypical data cells. This information can be visualized in a heatmap by using a color-coding of the weights. The appearance of the heatmap depends a lot on the chosen weight function, but the choice of the weight function would not essentially change a resulting ROC curve (this is valid at least for the weight 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.

Based on simulations using artificially contaminated real data, the performance of cell-rPLR was compared to a state-of-the art cellwise outlier detection algorithm, where it turned out that cell-rPLR was more accurate and at the same time 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 presence of cellwise contamination.

The cell-rPLR is implemented in the software environment R (R Core Team, 2018) 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 (Chang et al., 2018), 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 weighting 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 distiction is difficult with the current version of the algorithm. Moreover, the cell-rPLR method will be extended to biomarker identification for the multi-group case.

Acknowledgments

This research is supported by FWF (Austrian Science Fund) and the GACR (Czech Science Fundation) project number I 1910-N26 (15-34613L) and GACR project number 18-12204S.

CHAPTER 5

R implementation

This chapter describes two packages which have been written in the frame of this thesis, the packages `Biomarker` and `cellrPLR`. Both are written in the open source software R (R Core Team, 2018). The packages include the implementations of the algorithms rPLR from Chapter 3 and cell-rPLR from Chapter 4, respectively. They are freely available at `https://github.com/walachja`, and they can be installed after loading the `devtools` library, which enables the GitHub installation by the commands: `install_github("walachja/Biomarker")` and `install_github("walachja/cellrPLR")`.

The functionality of those two packages will be demonstrated on a simulated dataset accessible by the `gendata2` function from the cellrPLR package. The dataset has two groups, each consisting of 20 samples. There are 200 variables in the data, including 15 variables with discrimination power between the groups, thus biomarkers. The data are simulated based on the normal distribution (see Section 3.3), and contain additive as well as multiplicative noise. Furthermore, the artificial size-effect is also considered.

```
library(Biomarker)
library(cellrPLR)

data <- gendata2(n1=20, n2=20, v=200, peaks=1:15)
x <- data$X
g1 <- data$g1
g2 <- data$g2
```

```
## dimensionality
dim(x)
# [1]  40 200
```

5.1 Package Biomarker

The package `Biomarker` is an implementation of the robust pairwise log-ratio method (rPLR) described in Chapter 3. It includes the main function `biomarker` for the identification of the biomarkers in the data. The `biomarker` function has several parameters listed in Table 5.1. The corresponding rPLR method is based on the variation matrix, which can be estimated classically, using the empirical standard deviation, or robustly, with either the MAD or the τ estimator, see Section 3.2. The indexes of the observations belonging to group 1 and 2 must be defined. The final information if the variable is or is not identified as a biomarker depends on the predetermined cut-off value. The default cut-off corresponds to the 97.5 quantile of the standard normal distribution. Furthermore, `plot`, `print` and `summary` functions are available and can be applied on the resulting object. The code below shows how to run the biomarker function, and the plot results are shown in Figure 5.1. It is a plot of the V_j^* values (see Equation (3.5)) with the cut-off line. Each value above the cut-off line was identified as a biomarker. Thus, the first 15 variables were correctly identified as biomarkers.

```
res1 <- biomarker(x=x, type= 'tau', g1=g1, g2=g2, plot=TRUE,
   diag=TRUE)
res1
# Number of identified biomarkers:  15
# Positions of Identified biomarkers:  1 2 3 4 5 6 7 8 9 10 11 12 13
   14 15
```

5.2 Package cellrPLR

The cell-rPLR package can be used for two purposes. Firstly, the identification of the biomarkers is implemented in the function `cellrPLR_biom`. Secondly, the identification of cellwise outliers is implemented in the function `plot_cellheatmap` and `cell_shinyApp`.

Table 5.1: Arguments of `biomarker` function, their explanation and default value.

Function argument	Description	Default setting
`x`	data matrix	
`g1`	vector of locations of observations of gr. 1	
`g2`	vector of locations of observations of gr. 2	
`type`	type of variation matrix	`"tau"`
`cut`	cut-off value	$u_{0.975}$
`diag`	logical; compute outlier diagnostic	`TRUE`
`plot`	logical; plot results	`TRUE`
`diag.plot`	logical; plot outlier diagnostics	`FALSE`

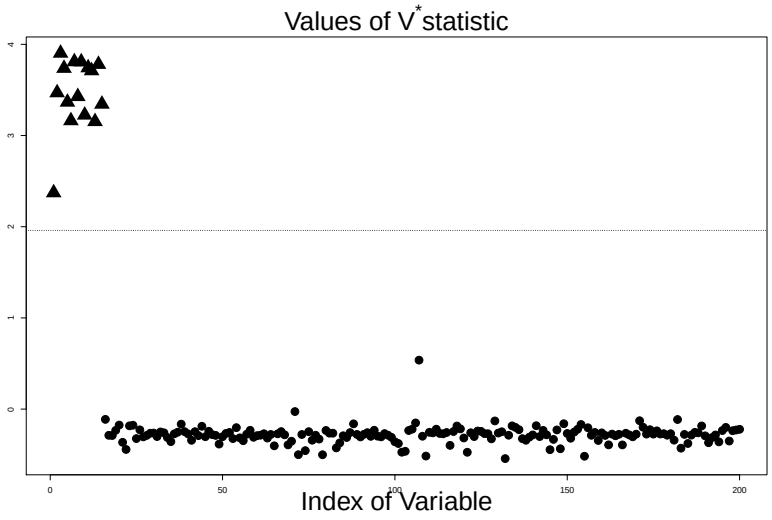


Figure 5.1: Resulting plot from function `biomarker`.

cellrPLR_biom

The input arguments of the `cellrPLR_biom` function are described in Table 5.2. The type of the weighting function used for the computation needs to be specified. Furthermore, the group structure defined by the parameters `g1` and `g2` must be included. The group which is used as a basis (see Section 4.2.5) can be specified by the parameter `mainGroup`, where the `"max"` value will automatically select the bigger of the groups as the basis for centering and scaling the data. The `permutation` parameter specifies if the permutation tests, as described in Section 4.2.5, should be employed. The additional parameters `B` and `p.alpha` refer to the number of the permutation repetitions and to the permutation test cut-off value, respectively. If the true biomarkers are known, they can be specified in the `biomarkers` parameter. Thus, the ranking positions of these biomarkers are

returned, together with some basis information such as True Positive and False Positive Rate.

The results are returned as a data frame containing the computed differences in the interval $[-1, 1]$, see Equation (4.16), the ranked list of the most discriminating variables, and possibly results of the permutation tests.

```
res2 <- cellrPLR_biom(data, type = 'biweight', mainGroup = 1, g1 =
   g1, g2 = g2, biomarker = 1:15,permutation = TRUE, B = 100,p.alpha
   = 0.95)
```

Table 5.2: The arguments of the `cellrPLR_biom` function, their explanation and default value.

Function argument	Description	Default setting
`data`	data matrix	
`type`	type of weighting function	`"biweight"`
`g1`	index vector for observations from group 1	
`g2`	index vector for observations from group 2	
`mainGroup`	group chosen as basis group	`"max"`
`biomarkers`	names of biomarker variables	
`permutation`	logical; perform permutation tests	`FALSE`
`B`	number of iterations for perm. tests	`1000`
`p.alpha`	cut-off p-value for permutation tests	`0.95`

plot_cellheatmap

The function `plot_cellheatmap` is the main function for the analysis of the identification of cellwise outliers. It creates a heatmap of cellwise outlier information. The function has the same parameters as `cellrPLR_biom`, except e.g. `permutation`, since the permutation tests cannot be employed in this scheme. Furthermore, the function returns the final weight value for each cell. Figure 5.2 presents the resulting heatmap applied on the simulated dataset. A red value corresponds to a "positive outlier", and to a value which is higher than expected. A blue value indicates a "negative outlier", with a value lower than expected. The same figure can also be plotted from the resulting object of the `cellrPLR_biom` function, using the `plot` function. Since group 1 was selected as the basis, most of the cells of the biomarker variables for the group 2 are marked as outlying.

```
plot_cellheatmap(data = data, type = 'biweight', mainGroup = 1, g1 =
    g1, g2 = g2, plotly = TRUE, grid = TRUE, title = 'Simulated
    Example')

# Gives the same result as:
plot(res2, plotly = TRUE, grid = TRUE, title = 'Simulated Example')
```

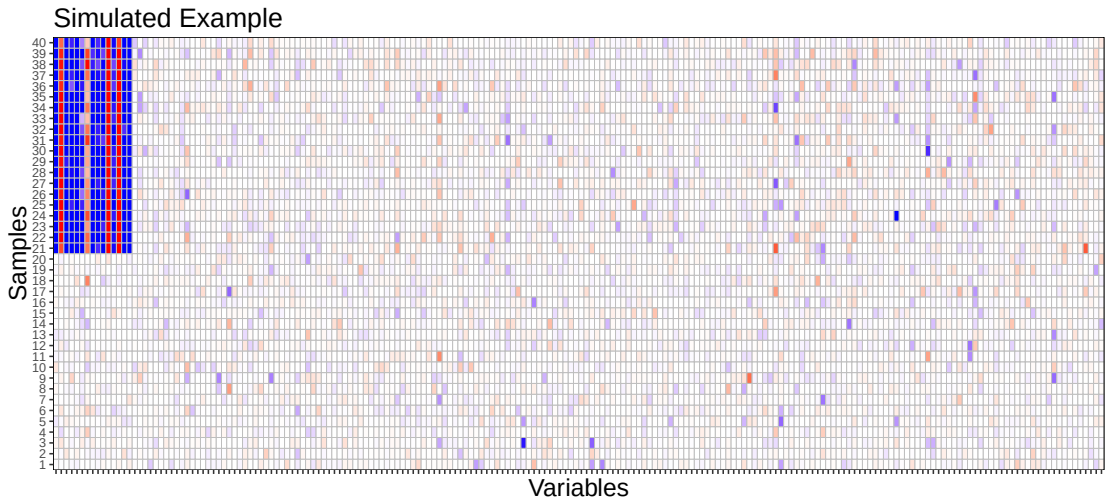


Figure 5.2: Resulting heatmap from the function `plot_cellheatmap`.

List of Figures

List of Tables

Bibliography

T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2009.

C. Agostinelli, A. Basu, P. Filzmoser, and D. Mukherjee. *Recent Advances in Robust Statistics: Theory and Applications*. Springer, New Delhi, India, 2016.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, UK, 1986.

J. Aitchison. Measures of location of compositional data sets. *Mathematical Geology*, 21 (7):787–790, 1989.

J. Aitchison and S. Shen. Logistic-normal distributions: Some properties and uses. *Biomatrika*, 67(2):173–185, 1980.

A. Alonso, S. Marsal, and A. Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3:23, 2015.

D. G. Altman and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *The Statistician*, 32(3):307–317, 1983.

E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1-2):237–260, 1998.

M. Åstrand. Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology*, 10(1):95–102, 2003.

A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

R. Bellio and L. Ventura. An introduction to robust estimation with R functions. In *Proceedings of 1st International Workshop on Robust statistics and R*, pages 1–57. Treviso: Department of Statistics, Ca'Foscari University ,Venezia, Italy, 2005.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

K. P. Bennett, U. Fayyad, and D. Geiger. Density-based indexing for approximate nearest-neighbor queries. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 233–243. ACM, New York, NY, USA, 1999.

K. Beyer, J. Goldstein, R. Ramakrishnan, e. C. Shaft, Uri", and P. Buneman. When is "nearest neighbor" meaningful? In *Database Theory — ICDT'99*, pages 217–235. Springer, Berlin, Heidelberg, Germany, 1999.

P. J. Bickel and K. A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.

B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

K. Burgess, N. Rankin, and S. Weidt. Chapter 10 - metabolomics. In S. Padmanabhan, editor, *Handbook of Pharmacogenomics and Stratified Medicine*, pages 181 – 205. Academic Press, San Diego, CA, USA, 2014.

V. Centner, D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna. Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry*, 68(21):3851–3858, 1996.

W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2018. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.1.0.

T. Chen, Y. Cao, Y. Zhang, J. Liu, Y. Bao, C. Wang, W. Jia, and A. Zhao. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-Based Complementary and Alternative Medicine*, 2013:11, 2013.

I. G. Chong and C. H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1): 103–112, 2005.

W. S. Cleveland and S. J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403): 596–610, 1988.

A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78 (7):2262–2267, 2006.

C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.

E. Dudley, M. Yousef, Y. Wang, and W. Griffiths. Targeted metabolomics and mass spectrometry. In *Advances in Protein Chemistry and Structural Biology*, volume 80, pages 45–83. Elsevier, Oxford, UK, 2010.

S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 10(1):111–139, 2002.

C. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

E. Edgington and P. Onghena. *Randomization tests*. Chapman and Hall/CRC, New York, USA, 2007.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003.

L. Eriksson. *Introduction to multi-and megavariate data analysis using projection methods (PCA & PLS)*. Umetrics, Umeå, Schweden, 1999.

S. Favilla, C. Durante, M. L. Vigni, and M. Cocchi. Assessing feature relevance in NPLS models by VIP. *Chemometrics and Intelligent Laboratory Systems*, 129:76–86, 2013.

A. D. Fernandes, J. M. Macklaim, T. G. Linn, G. Reid, and G. B. Gloor. Anova-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One*, 8 (7):e67019, 2013.

P. Filzmoser and B. Walczak. What can go wrong at the data normalization step for identification of biomarkers? *Journal of Chromatography A*, 1362:194–205, 2014.

P. Filzmoser, R. Maronna, and M. Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

P. Filzmoser, S. Serneels, R. Maronna, and P. J. Van Espen. Robust multivariate methods in chemometrics. In B. Walczak, R. Ferre, and S. Brown, editors, *Comprehensive Chemometrics (vol. 3)*, page 681–722. Elsevier, Oxford, UK, 2009.

P. Filzmoser, K. Hron, and M. Templ. *Applied Compositional Data Analysis: With Worked Examples in R*. Springer, Cham, Germany, 2018.

E. Fišerová and K. Hron. On the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43(4):455, 2011.

R. A. Fisher. *The Design of Experiments.* Oliver & Boyd, Edinburgh and London, UK, 1935.

P. Franceschi, D. Masuero, U. Vrhovsek, F. Mattivi, and R. Wehrens. A benchmark spike-in data set for biomarker identification in metabolomics. *Journal of Chemometrics*, 26 (1-2):16–24, 2012.

A. H. Garde, Å. M. Hansen, J. Kristiansen, and L. E. Knudsen. Comparison of uncertainties related to standardization of urine samples with volume and creatinine concentration. *Annals of Occupational Hygiene*, 48(2):171–179, 2004.

A. Gardlo. *Multidimensional statistical methods for analysis of human metabolome*. PhD dissertation, Palacký University Olomouc, Czech Republic, 2016.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

P. Giraudeau, I. Tea, G. S. Remaud, and S. Akoka. Reference and normalization methods: Essential tools for the intercomparison of NMR spectra. *Journal of Pharmaceutical and Biomedical Analysis*, 93(Supplement C):3–16, 2014.

G. B. Gloor and G. Reid. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8):692–703, 2016.

R. Gosselin, D. Rodrigue, and C. Duchesne. A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemometrics and Intelligent Laboratory Systems*, 100(1):12–21, 2010.

P. S. Gromski, Y. Xu, K. A. Hollywood, M. L. Turner, and R. Goodacre. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics*, 11(3): 684–695, 2015.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

F. R. Hampel, P. J. Rousseeuw, and E. Ronchetti. The change-of-variance curve and optimal redescending m-estimators. *Journal of the American Statistical Association*, 76(375):643–648, 1981.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, New York, NY, USA, 1986.

F. E. Harrell. *Regression modeling strategies.* Springer, Cham, Germany, 2014.

Z. M. Hira and D. F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015:13, 2015.

J. Hochrein, H. U. Zacharias, F. Taruttis, C. Samol, J. C. Engelmann, R. Spang, P. J. Oefner, and W. Gronwald. Data normalization of 1H NMR metabolite fingerprinting data sets in the presence of unbalanced metabolite regulation. *Journal of Proteome Research*, 14(8):3217–3228, 2015.

I. Hoffmann, P. Filzmoser, S. Serneels, and K. Varmuza. Sparse and robust PLS for binary classification. *Journal of Chemometrics*, 30(4):153–162, 2016.

J. H. Holland. Genetic algorithms. *Scientific American*, 267(1):66–73, 1992.

P. J. Huber. Robust statistics. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, Berlin, Heidelberg, Germany, 2011.

P. J. Huber and E. Ronchetti. *Robust Statistics, Series in Probability and Mathematical Statistics.* John Wiley, New York, NY, USA, 1981.

W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl_1):S96–S104, 2002.

M. Hubert, P. J. Rousseeuw, and S. Verboven. A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1-2):101–111, 2002.

M. Hubert, P. J. Rousseeuw, and K. Vanden Branden. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

J. E. Jackson. *A User's Guide to Principal Components.* John Wiley & Sons, New York, USA, 2005.

H. Janečková, K. Hron, P. Wojtowicz, E. Hlídková, A. Barešová, D. Friedeckỳ, L. Žídková, P. Hornik, D. Behúlová, D. Procházková, et al. Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders. *Journal of Chromatography A*, 1226:11–17, 2012.

R. S. Jansen, R. Addie, R. Merkx, A. Fish, S. Mahakena, O. B. Bleijerveld, M. Altelaar, L. Ijlst, R. J. Wanders, P. Borst, et al. N-lactoyl-amino acids are ubiquitous metabolites that originate from CNDP2-mediated reverse proteolysis of lactate and amino acids. *Proceedings of the National Academy of Sciences*, 112(21):6601–6606, 2015.

I. Jolliffe. Principal component analysis. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, Berlin, Heidelberg, Germany, 2011.

M. Katajamaa and M. Orešič. Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, 1158(1-2):318–328, 2007.

O. Kempthorne. *The Design and Analysis of Experiments.* John Wiley & Sons, New York, NY, USA, 1952.

H. C. Keun, T. M. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon, and J. K. Nicholson. Improved analysis of multivariate data by variable

stability scaling: application to nmr-based metabolic profiling. *Analytica Chimica Acta*, 490(1):265–276, 2003.

J. Kittler. Feature selection and extraction. In Young and Fu, editors, *Handbook of Pattern Recognition and Image Processing*. Academic Press, New York, NY, USA, 1986.

S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, and W. Gronwald. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8(1):146–160, 2012.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

H. Kubinyi. *3D QSAR in drug design: Volume 1: Theory Methods and Applications*. Springer Science & Business Media, The Netherlands, 1994.

C. Kuhl, R. Tautenhahn, C. Böttcher, T. Larson, and S. Neumann. Camera: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2012.

O. M. Kvalheim. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *Journal of Chemometrics*, 24(7-8):496–504, 2009.

O. M. Kvalheim, F. Brakstad, and Y. Liang. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry*, 66(1):43–51, 1994.

T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. Embedded methods. In I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, editors, *Feature extraction*, pages 137–165. Springer, Berlin, Heidelberg, Germany, 2006.

C. Lambeth, W. Gladstone, and R. Stonecypher. Statistical efficiency of row and noncontiguous family plots in genetic tests of loblolly pine. *Silvae Genet*, 32:24–28, 1983.

A. Lazraq, R. Cléroux, and J. P. Gauchi. Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, 66 (2):117–126, 2003.

B. Li, J. Tang, Q. Yang, X. Cui, S. Li, S. Chen, Q. Cao, W. Xue, N. Chen, and F. Zhu. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Scientific Reports*, 6, 2016.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Journal of the American Statistical Association*, 80(391):759–766, 1985.

Y.-Z. Liang and K.-T. Fang. Robust multivariate calibration algorithm based on least median of squares and sequential number theory optimization method. *Analyst*, 121 (8):1025–1029, 1996.

Y.-Z. Liang and O. M. Kvalheim. Robust methods for multivariate analysis—a tutorial review. *Chemometrics and Intelligent Laboratory Systems*, 32(1):1–10, 1996.

J. Lindon, E. Holmes, and J. Nicholson. So what's the deal with metabonomics? *Analytical Chemistry*, 75(17):384A–391A, 2003.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

R. Maronna, R. D. Martin, and V. Yohai. *Robust statistics.* John Wiley & Sons, Chichester, UK, 2006.

R. A. Maronna and R. H. Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317, 2002.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust Statistics: Theory and Methods (with R).* Wiley, Chichester, UK, 2019.

J. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9):2688–2704, 2012.

T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, 2012.

M. J. Miller, A. D. Kennedy, A. D. Eckhart, L. C. Burrage, J. E. Wulff, L. A. Miller, M. V. Milburn, J. A. Ryals, A. L. Beaudet, Q. Sun, et al. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *Journal of Inherited Metabolic Disease*, 38(6):1029–1039, 2015.

M. Monteiro, M. Carvalho, M. Bastos, and P. Guedes de Pinho. Metabolomic analysis for biomarker discovery: Advances and challenges. *Current Medicinal Chemistry*, 20 (2):257–271, 2013.

L. Najdekr, A. Gardlo, L. Mádrová, D. Friedeckỳ, H. Janečková, E. S. Correa, and T. Goodacre, R.and Adam. Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-coa dehydrogenase deficiency. *Talanta*, 139:62–66, 2015.

J. K. Nicholson, J. Connelly, J. C. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 1(2): 153, 2002.

S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16(9):373–378, 1998.

V. Öllerer, A. Alfons, and C. Croux. The shooting S-estimator for robust regression. *Computational Statistics*, 31(3):829–844, 2016.

G. J. Patti, O. Yanes, and G. Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263, 2012.

V. Pawlowsky-Glahn and A. Buccianti. *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Chichester, UK, 2011.

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

M. S. Pepe, R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget, and Y. Yasui. Phases of biomarker development for early detection of cancer. *JNCI: Journal of the National Cancer Institute*, 93(14):1054–1061, 2001.

M. Pérez-Enciso and M. Tenenhaus. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, 112(5-6):581–592, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

P. D. Rainville, G. Theodoridis, R. S. Plumb, and I. D. Wilson. Advances in liquid chromatography coupled to mass spectrometry for metabolic phenotyping. *TrAC Trends in Analytical Chemistry*, 61:181–191, 2014.

T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik, and O. M. Kvalheim. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems*, 95(1):35–48, 2009a.

T. Rajalahti, R. Arneberg, A. C. Kroksveen, M. Berle, K.-M. Myhr, and O. M. Kvalheim. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical Chemistry*, 81(7):2581–2590, 2009b.

J. Raymaekers, P. Rousseeuw, and W. Van den Bossche. *cellWise: Analyzing Data with Cellwise Outliers*, 2018. URL `https://CRAN.R-project.org/package=cellWise`. R package version 2.0.10.

H. W. Ressom, J. F. Xiao, L. Tuli, R. S. Varghese, B. Zhou, T. Tsai, M. R. Ranjbar, Y. Zhao, J. Wang, C. Di Poto, et al. Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis. *Analytica Chimica Acta*, 743:90–100, 2012.

U. Roessner and D. A. Dias. *Metabolomics tools for natural product discovery*. Springer, 2016.

P. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

P. J. Rousseeuw and W. V. D. Bossche. Detecting deviating data cells. *Technometrics*, 60(2):135–145, 2018.

D. B. Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

E. Saccenti. Correlation patterns in experimental data are affected by normalization procedures: Consequences for data analysis and network inference. *Journal of Proteome Research*, 16(2):619–634, 2017.

E. Saccenti, H. C. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. Hendriks. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3):361–374, 2014.

Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

F. Schroeder. *Cost Sensitive Screening Methods for Binary Classification*. PhD dissertation, TU Wien, Austria, 2018.

J. Scott Long. Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques in the Social Sciences*, 7, 1997.

S. Serneels. and I. Hoffmann. *SPRM: Sparse and Non-Sparse Partial Robust M Regression and Classification*, 2016. URL `https://CRAN.R-project.org/package=sprm`. R package version 1.2.2.

S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen. Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 79(1-2):55–64, 2005.

S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen. The partial robust M-approach. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, pages 230–237. Springer, Berlin, Heidelberg, Germany, 2006.

Y. I. Shurubor, U. Paolucci, B. F. Krasnikov, W. R. Matson, and B. S. Kristal. Analytical precision, biological variation, and mathematical normalization in high data density metabolomics. *Metabolomics*, 1(1):75–85, 2005.

D. A. Skoog, F. J. Holler, and S. R. Crouch. *Principles of Instrumental Analysis*. Cengage learning, Boston, USA, 2017.

A. K. Smilde, M. J. van der W., S. Bijlsma, B. J. van der Werff-van der Vat, and R. H. Jellema. Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, 77(20):6729–6736, 2005.

C. Smith, E. Want, G. O'Maille, R. Abagyan, and R. Siuzdak. Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry*, 78(3):779–787, 2006. doi: 10.1021/ac051437y.

L. Ståhle and S. Wold. Partial least squares analysis with cross-validation for the two-class problem: A monte carlo study. *Journal of Chemometrics*, 1(3):185–196, 1987.

K. Strimbu and J. A. Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

M. Sysi-Aho, M. Katajamaa, L. Yetukuri, and M. Orešič. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, 8(1):93, 2007.

M. Templ, K. Hron, P. Filzmoser, and A. Gardlo. Imputation of rounded zeros for high-dimensional compositional data. *Chemometrics and Intelligent Laboratory Systems*, 155:183–190, 2016.

M. Templ, K. Hron, and P. Filzmoser. *robCompositions: an R-package for robust statistical analysis of compositional data*, 2017. URL `https://CRAN.R-project.org/package=robCompositions`.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

V. Todorov and A. M. Pires. Comparative performance of several robust linear discriminant analysis methods. *REVSTAT Statistical Journal*, 5:63–83, 2007.

Y. Truong, X. Lin, and C. Beecher. Learning a complex metabolomic dataset using random forests and support vector machines. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 835–840. ACM, 2004.

Y. Tsuchiya, Y. Takahashi, T. Jindo, K. Furuhama, and K. T. Suzuki. Comprehensive evaluation of canine renal papillary necrosis induced by nefiracetam, a neurotransmission enhancer. *European Journal of Pharmacology*, 475(1):119–128, 2003.

J. Václavík, K. L. Coene, I. Vrobel, L. Najdekr, D. Friedeckỳ, R. Karlíková, L. Mádrová, A. Petsalo, U. F. Engelke, A. van Wegberg, et al. Structural elucidation of novel

biomarkers of known metabolic disorders based on multistage fragmentation mass spectra. *Journal of Inherited Metabolic Disease*, 41(3):407–414, 2018.

R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(1):142, Jun 2006.

M. J. Van Der Werf, R. H. Jellema, and T. Hankemeier. Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *Journal of Industrial Microbiology and Biotechnology*, 32(6):234–252, 2005.

S. S. Waikar, V. S. Sabbisetti, and J. V. Bonventre. Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. *Kidney International*, 78(5): 486–494, 2010.

J. Walach, P. Filzmoser, K. Hron, B. Walczak, and L. Najdekr. Robust biomarker identification in a two-class problem based on pairwise log-ratios. *Chemometrics and Intelligent Laboratory Systems*, 171:277–285, 2017.

J. Walach, P. Filzmoser, and K. Hron. Data normalization and scaling: Consequences for the analysis in omics scienc. In J. Jaumot, C. Bedia, and R. Tauler, editors, *Data Analysis for Omics Sciences: Methods and Applications*, pages 65–196. Elsevier, Amsterdam, The Netherlands, 2018.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3): 426–482, 1943.

J. Wang, T. T. Christison, K. Misuno, L. Lopez, A. F. Huhmer, Y. Huang, and S. Hu. Metabolomic profiling of anionic metabolites in head and neck cancer cells by capillary ion chromatography with orbitrap mass spectrometry. *Analytical Chemistry*, 86(10): 5116–5124, 2014.

B. M. Warrack, S. Hnatyshyn, K.-H. Ott, M. D. Reily, M. Sanders, H. Zhang, and D. M. Drexler. Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B*, 877(5-6):547–552, 2009.

B.-J. M. Webb-Robertson, D. F. Lowry, K. H. Jarman, S. J. Harbo, Q. R. Meng, A. F. Fuciarelli, J. G. Pounds, and K. M. Lee. A study of spectral integration and normalization in nmr-based metabonomic analyses. *Journal of Pharmaceutical and Biomedical Analysis*, 39(3):830–836, 2005.

R. Wehrens, P. Franceschi, U. Vrhovsek, and F. Mattivi. Stability-based biomarker selection. *Analytica Chimica Acta*, 705(1-2):15–23, 2011.

B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362, 1938.

H. Wold. Path models with latent variables: The NIPALS approach. In H. Blalock, A. Aganbegian, F. Borodkin, R. Boudon, and V. Capecchi, editors, *Quantitative Sociology*, International Perspectives on Mathematical and Statistical Modeling, pages 307 – 357. Academic Press, London, UK, 1975.

S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the pls method. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, pages 286–293. Springer, Berlin, Heidelberg, Germany, 1983.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

S. Wold, E. Johansson, and M. Cocchi. PLS—partial least squares projections to latent structures. 3D QSAR. *Drug Design*, 1:523–550, 1993.

S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielser, H.-H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3(9):1–16, 2002.

R. Wu, X. Zhao, Z. Wang, M. Zhou, and Q. Chen. Novel molecular events in oral carcinogenesis via integrative approaches. *Journal of Dental Research*, 90(5):561–572, 2011.

P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya. A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. *BMC Bioinformatics*, 11(1):S5, 2010.

V. J. Yohai and R. H. Zamar. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83(402):406–413, 1988.

P. Zerzucha and B. Walczak. Again about partial least squares and feature selection. *Chemometrics and Intelligent Laboratory Systems*, 115:9–17, 2012.

Curriculum Vitae

Jan Walach

Contact address
TU Wien
Institute of Statistics and Mathematical Methods in Economics
Research Group Computational Statistics (CSTAT)
Wiedner Hauptstraße 8-10

A-1040 Vienna, Austria
Email: jan.walach@tuwien.ac.at, walach.jan@gmail.com

Phone: +43 67761747188, +420 737192227

Education

---

since 09/2015	**TU Wien, Vienna (Austria)**
	Ph.D. candidate in Statistics
2013 – 2015	**Palacký University Olomouc (Czech Republic)**
	Masters in Applications of Mathematics in Economy
	Thesis: *Image processing methods*
2010 – 2013	**Palacký University Olomouc (Czech Republic)**
	Bachalor in Mathematics - Economics with focus on banking
	Thesis: *Evaluation of vowel quality by means of the formant theory and Fourier Analysis*
2013 – 2014	**University of Trento (Italy)**
	Erasmus program

Working experience

since 09/2015	**TU Wien, Vienna (Austria)**
	Project assistant (FWF) for *Statistics in metabolomics for biomarker research in medicine*
since 03/2018	**AutoCont CZ. a.s. (Czech Republic)**
	Data Scientist

Publications

J. Walach, P. Filzmoser, K. Hron, B. Walczak and L. Najdekr. Robust biomarker identification in a two-class problem based on pairwise log-ratios. Chemometrics and Intelligent Laboratory Systems, 171, pp. 277-285, 2017

J. Walach, P. Filzmoser, and K. Hron. Data normalization and scaling: Consequences for the analysis in omics sciences. In: J. Jaumot, C. Bedia, and R. Tauler (eds.) Comprehensive Analytical Chemistry. Data Analysis for Omics Sciences: Methods and Applications. Elsevier, Amsterdam, The Netherlands, pp. 165-196, 2018.

J. Walach, P. Filzmoser and Š. Kouřil. Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log-ratios, submitted, 2018

J. Tobin, J. Walach, D. de Beer, P. J. Williams, P. Filzmoser and B. Walczak. Untargeted anlysis of chromatographic data for green and fermented rooibos: problem with size effect removal. Journal of Chromatography A, 1525, pp. 109-115, 2017.

K. Varmuza, P. Filzmoser, I. Hoffmann, J. Walach et. al. Significance of variables for discrimination: Applied to the search of organic ions in mass spectra measured on cometary particles. Journal of Chemometrics; 32(4):e3001, 2018