Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

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

Jan Walach [a,*], Peter Filzmoser [a], Karel Hron [b], Beata Walczak [c], Lukáš Najdekr [d,e]

[a] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
[b] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacky University, 17.listopadu 12, 77146 Olomouc, Czech Republic
[c] Department of Analytical Chemistry, Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland
[d] Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacký University in Olomouc, Hněvotínská 5, 775 15 Olomouc, Czech Republic
[e] University Hospital Olomouc, I.P. Pavlova 185/6, 779 00 Olomouc, Czech Republic

ARTICLE INFO

Keywords:
Biomarker
Log-ratio methodology
Size effect
Robust method
rPLR

ABSTRACT

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.

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) [1,2]. 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 [3].

An alternative is probabilistic quotient normalization (PQN) [4]. 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 $x_{\mathrm{ref},j}$ for each variable, $s_i = \mathrm{median}(x_{i1}/x_{ref,1},\ldots,x_{id}/x_{ref,d})$. The reference values are the column medians or means of $\mathbf{X}$ [4]. 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 [3], 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 [5]. It turned out that PQN

* Corresponding author. TU Wien, Wiedner Hauptstraße 8-10/105, A-1040 Wien, Austria.
E-mail addresses: jan.walach@tuwien.ac.at (J. Walach), peter.filzmoser@tuwien.ac.at (P. Filzmoser), karel.hron@upol.cz (K. Hron), beata.walczak@us.edu.pl (B. Walczak), lukas.najdekr@upol.cz (L. Najdekr).

https://doi.org/10.1016/j.chemolab.2017.09.003
Received 17 January 2017; Received in revised form 6 September 2017; Accepted 8 September 2017
Available online xxxx

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 $\mathrm{x} = (x_1, \ldots, x_d)$. In presence of a scaling constant we observe $s{\cdot}\mathrm{x} = (s{\cdot}x_1, \ldots, s{\cdot}x_d)$. However, the ratios between any two variables of the true signal, $x_j\,/x_k$, carries the same information as the corresponding ratios of $s{\cdot}\mathrm{x}$, 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 Ref. [6], ratios are not easy to deal with, because their variances are non-symmetrical, since $var(x_j/x_k) \neq var(x_k/x_j)$. This was solved by using logarithms of ratios, so called log-ratios, which meet the property of symmetry, since $var(\ln(x_j/x_k)) = var(\ln(x_k/x_j))$. Log-ratios are used in the field of compositional data analysis [5].

The main goal of this study is to present a new method for biomarker identification based on robust Pair-wise Log-Ratios: rPLR (Section 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 [7,8]. 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. Section 4 presents new ways of outlier diagnostics, which also lead to interesting findings in a real data example in Section 5. The final Section 6 provides concluding remarks.

2. Method rPLR

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

2.1. Variation matrix

The proposed method builds on the variation matrix $\mathbf{T}$ [9,10], with the elements t_{jk} defined as:

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

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 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

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

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

2.2. Test statistic

For biomarker identification, the following statistic V_j is proposed,

$$V_j = \sum_{k=1}^{d} \frac{n_1{\cdot}\sqrt{t_{jk}^{(1)}} + n_2{\cdot}\sqrt{t_{jk}^{(2)}}}{(n_1 + n_2){\cdot}\sqrt{t_{jk}}}, \qquad \text{for } j = 1, \ldots, d. \tag{3}$$

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) 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) is, the less similar the groups are with respect to this jth variable.

Note that in Eq. (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

$$V_j = \sum_{k=1}^{d} \frac{\sqrt{t_{jk}^{(1)}} + \sqrt{t_{jk}^{(2)}}}{2{\cdot}\sqrt{t_{jk}}}, \qquad \text{for } j = 1, \ldots, d. \tag{4}$$

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 and 4 is used in order keep the values of V_j more symmetric, hence closer to normality.

We consider a normalized version

$$V_j^* = -\frac{V_j - \overline{V}}{s_V}, \qquad for \quad j = 1, \ldots, d, \tag{5}$$

with the arithmetic mean

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

and the empirical standard deviation

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

Because of the minus sign in (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. Refs. [11,12]) as an alternative. Randomization tests do not assume normality or any other distribution of the data, but they are computationally much