Journal of Chromatography A, 1362 (2014) 194–205

Contents lists available at ScienceDirect

Journal of Chromatography A

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

What can go wrong at the data normalization step for identification of biomarkers?

P. Filzmoser[a], B. Walczak[b,*]

[a] Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria
[b] Department of Analytical Chemistry, Institute of Chemistry, University of Silesia, Katowice, Poland

ARTICLE INFO

Article history:
Received 20 March 2014
Received in revised form 12 August 2014
Accepted 14 August 2014
Available online 21 August 2014

Keywords:
Log-ratio methodology
Size effect
Shape effect
Fingerprints
Compositional data analysis.

ABSTRACT

Our study focuses on the removal of the so-called size effect, related to a different sample volume and/or concentration. This effect is associated with many types of instrumental signals, particularly with those originating from HPLC-DAD, LC-MS, and UPLC-MS. These signals do not carry any absolute information about the sample components. If the data comparison has to be performed based on sample fingerprints, then the size effect is undesired, and the shape effect is of main interest. With "shape", we refer to data information which is contained in the ratios between the variables. So far, different normalization methods have been applied to the removal of size effect. In our study, the performance of popular normalization methods is compared with those of the CODA (Compositional Data Analysis) methods, relying on log-ratio transformations, and the performance is evaluated through the prism of proper identification of biomarkers.

1. Introduction

Hyphenated chromatographic techniques (such as, e.g., HPLC-DAD, LC-MS, and UPLC-MS) have nowadays become a standard analytical tool for studying complex biological samples. Generated chromatograms (fingerprints) require extensive pre-processing, prior to comparative statistical analyses. The pre-processing step consists of signals enhancement (de-noising and background elimination), warping and removal of the so-called 'size effect', associated with a different sample volume and/or sample concentration. Due to the 'size effect', the true signal is unobservable, but what is observed is a signal that is multiplied by a constant, and the constants in general differ for different signals. Our study focuses on the 'size effect' removal, i.e., on data normalization. Although there are many methods which can be applied at this step of data analysis, the choice of a right method is not obvious and it depends on data characteristics and the problem at hand [1,2]. It can happen that the methods which are considered as alternative approaches lead to different results and different conclusions. As demonstrated in [3], the majority of justified (and often applied) pre-processing methods lead to deterioration of model performance. In our study, the influence of data normalization methods on the 'biomarker' identification is investigated.

2. Elimination of size effect

Nearly any step in a sample preparation and analysis can contribute to the variation of peak areas. This type of variation (experimental error) differs, however, from the so-called biological error, representing natural variability of the studied samples [4]. In the studies of natural samples, the biological error usually dominates over the experimental error, part of which can be controlled or eliminated (e.g., instrumental noise). Due to the experimental and natural variability, chromatograms of the studied samples usually differ in size and shape. They do not carry any absolute information about the concentration of sample components. In the case of targeted analysis, absolute concentration value(s) can obviously be calculated based on the calibration with the known standards. If data comparison has to be performed based on sample fingerprints, then the size effect is undesired, and the shape effect is of main interest. With "shape", we refer to data information which is contained in the ratios between the variables. In order to eliminate size effect, different types of the data normalization methods are applied.

Unfortunately, quite often chromatographic signals are normalized to the total sum (total sum normalization, TSN) [1]. This normalization causes that the sum of the data values for each observation equals 1 (or remains constant). TSN is not always justified and it can lead toward the wrong conclusions. It introduces the so-called closure (defined as the data where the observations add to a constant sum). With closed data, an increase in the concentration

* Corresponding author.
E-mail address: beata.walczak@us.edu.pl (B. Walczak).

http://dx.doi.org/10.1016/j.chroma.2014.08.050
0021-9673/

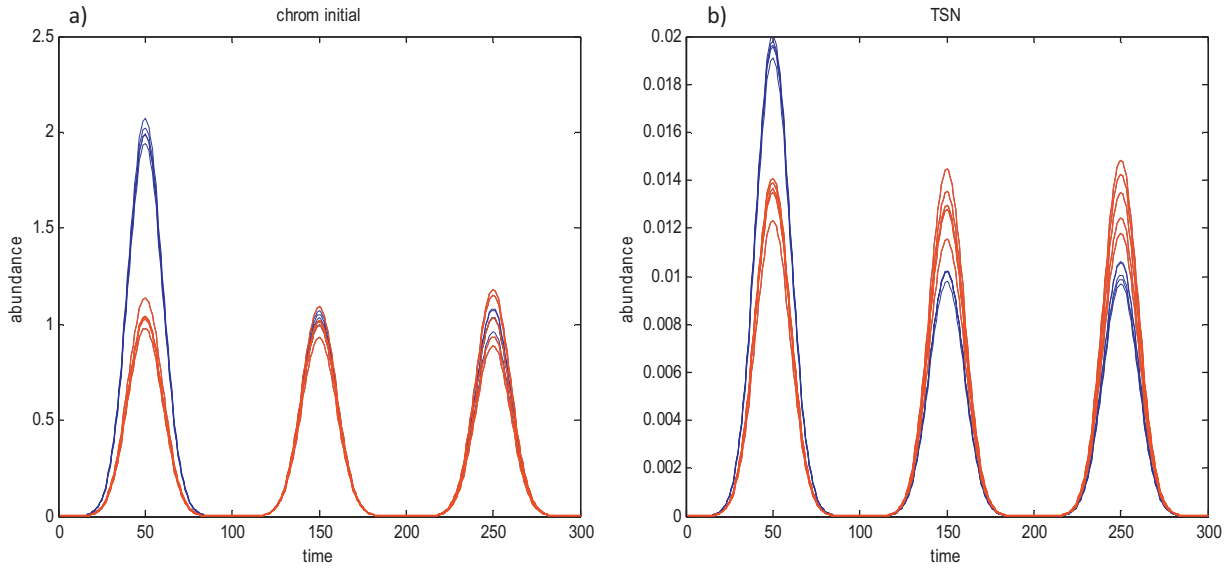

Fig. 1. (a) Initial data set containing samples belonging to the two classes, and (b) the same data set after normalization to the total sum.

of one component requires a decrease in the concentrations of the remaining components. The effect of applying TSN for identification of discriminatory peaks is illustrated in Fig. 1.

Initially, samples from class 1 and class 2 differ due to component #1 (peak #1). After normalization to the total sum, this difference is distributed along the entire signals. Ultimately, this dramatic effect is going to influence final conclusions concerning identification of the components differentiating the studied samples ('biomarkers').

The list of the normalization methods applied to the LC-DAD or LC-MS signals is quite long. One can find normalization to the standard (single or multiple, internal or external), normalization to the highest peak, normalization to a constant Euclidean norm or median, probabilistic quotient normalization, normalization by the maximum likelihood method, etc. (e.g., see [1,5–8]). However, all these methods have certain limitations and often lead to different results.

2.1. Compositional data analysis

Problems associated with the size effect could be completely eliminated, if—instead of working with the original variables—we worked with the ratios thereof. This is easy to notice, if we assume that the signal $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ with n variables (components) cannot be directly observed, but only the signal $\boldsymbol{x}s = [x_1 s, \ldots, x_n s]$, which is a multiple of the unobserved signal with a constant s. Then the ratio between any two variables of the unobserved signal, x_i/x_j, is the same as the ratio of the observed values, $(x_i s)/(x_j s) = x_i/x_j$. Thus, the relevant information carried by chromatographic signals (fingerprints) is contained in the ratios between the variables. Here, compositional data analysis, CODA, enters our story, since its aim is to focus on the analysis of the (log-)ratios. This analysis has already found numerous applications in different fields of sciences, mainly in geology and geochemistry (e.g., see [9]), yet so far, its applications in chemistry have been very limited. The main reason that compositional data analysis has not been applied, e.g., in chromatography, is that compositional data are often defined as data consisting of vectors whose components make a proportion or percentage of a certain entity, and the sum thereof is constrained in order to remain constant [9]. According to this definition, chromatographic data are not compositional data. For instance, in the case of LC-DAD, the observed peaks represent only these sample components which absorb in the DAD detector range (i.e., in most cases, not all sample components), and absorption (as well as the peak area) depends on the component concentration and on its molar absorptivity. Two components of the same chemical concentration can be represented by the two peaks of different areas, due to the difference in their molecular absorptivities. A k-fold increase in the concentration of component A does not cause a k-fold decrease in the concentration of component B, so their sum is not constrained. However, by 'compositional data', we also understand that the size effect (multiplication of the signal with a constant) is irrelevant, or in other words, compositional data are the data which carry relative information only. This definition is much broader than the first one and it includes such data which do not sum to a constant. As stated by Aitchison [10,11], 'to acknowledge the fact that information is relative, any reasonable statement about a composition has to be in terms of ratios of components' (in our case, in terms of the variables, or peak areas ratios). Ratios are size-irrelevant, but unfortunately, ratios are not nice to deal with due to the asymmetry of their variance ($\mathrm{var}(\boldsymbol{x}_1/\boldsymbol{x}_2)$ is not equal to $\mathrm{var}(\boldsymbol{x}_2/\boldsymbol{x}_1)$). This undesired property disappears, when log-ratios are considered instead, because $\mathrm{var}(\log(\boldsymbol{x}_1/\boldsymbol{x}_2)) = \mathrm{var}(\log(\boldsymbol{x}_2/\boldsymbol{x}_1))$. The log-ratio transformation is a basic transformation in compositional data analysis. It was proposed as a solution for compositional data, possibly represented with a constant sum constraint, for which the geometry is not the Euclidean geometry. The log-ratio transformation removes the problem of a constrained sample space. It means that the closure constraint is not important any longer, because the ratios of the components of the original data vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ and the ratios of the components of the same vector after normalization to the total sum are the same, and the data after the log-ratio transformation can be analyzed by a classical multivariate method. As stated in [12], 'Compositions are better thought of as equivalence classes of vectors with positive components: two of these vectors are equivalent if their components are proportional. A further step is that components of compositions do not need to add to a constant' [13,14].

While working with chromatographic data, we do not have a problem with the data closure (data are not in the Simplex sample space), but we have problems associated with the undesired 'size effect'. As all log-ratio transformations are scale invariant, i.e., $[x_1, \ldots, x_n]$ and $[sx_1, \ldots, sx_n]$ contain essentially the same information, for any non-zero number s, the CODA approach seems to