Subspace discriminant index to expedite exploration of multi-class omics data

(Article begins on next page)

18 April 2024

Subspace Discriminant Index to Expedite Exploration of Multi-Class Omics Data

Sara Tortorella

Molecular Horizon, Via Montelino 30, 06084, Bettona, Perugia, Italy

Maurizio Servili

Department of Agricultural, Food and Environmental Sciences-DSA3, University of Perugia, 06126, Perugia, Italy

Tullia Gallina Toschi

Department of Agricultural and Food Science, University of Bologna, 40127, Bologna, Italy

Gabriele Cruciani

Department of Chemistry, Biology and Biotechnology, University of Perugia, 06123 Perugia, Italy

José Camacho

Departamento de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, 18071, Granada, Spain

Abstract

Omics datasets, comprehensively characterizing biological samples at a molecular level, are continuously increasing in both complexity and dimensionality. In this scenario, there is a need for tools to improve data interpretability, expediting the process of extracting relevant biochemical information. Here we introduce the subspace discriminant index (SDI) for multi-component models, which points to the most promising components to explore pre-defined groups of observations, and can also be used to compare several modeling variants in terms of discriminative power. The versatility and the efficiency of the proposed index is demonstrated in three real world *omics* case studies, from pharmaceutical and food research applications, including a highly complex multi-class problem. The SDI is especially useful during the initial exploration of a data set, in order to make informed decisions on, e.g., pre-processing or modeling variants for fur-

Email addresses: `sara@molhorizon.it`; 0000-0001-9691-8323 (Sara Tortorella), `maurizio.servili@unipg.it`; 0000-0002-8726-2075 (Maurizio Servili), `tullia.gallinatoschi@unibo.it`; 0000-0001-7241-2280 (Tullia Gallina Toschi), `gabri@chemiome.chm.unipg.it`; 0000-0002-4162-8692 (Gabriele Cruciani), `josecamacho@ugr.es`; 0000-0001-9804-8122 (José Camacho)

ther analysis. By boosting the interpretation capabilities, the SDI represents a significant addition to the chemometrics tool-set.

1. Introduction

The last two decades have been characterized by an exceptional technological evolution in the fields of analytical chemistry, biology and biotechnology. This catalyzed the rise of the so called *omics sciences*, namely genomics, transcriptomics, proteomics, metabolomics, aiming at comprehensively identifying genes, mRNA, proteins, or metabolites in biological samples, respectively. While differing on how samples are analyzed to address the different scientific questions, a common feature of all the *omics* datasets is that they are vast, extremely complex and hard to inspect and rationalize without appropriate tools [39]. For instance, lipidomics, a subfield of metabolomics, involves the identification and quantitation, usually by means of liquid chromatography mass spectrometry (LC-MS) [23, 24, 15, 21], of thousands of lipid molecular species within dozens, even hundreds of biological samples. In such a complex scenario, proper tools for advanced, still interpretable, data processing and mining are essential to fully elucidate biological meanings. As a consequence, the parallel major advance in the *omics* era has been the increased importance of bioinformatic and chemometric techniques used to mine such data [39, 43, 3, 36].

In this context, multivariate analysis has proven valuable in a high number of applications, e.g. [20, 34, 45, 35, 31, 1, 29, 42, 28, 32]. We can distinguish two settings for the application of analysis tools: the unsupervised and the supervised settings.

In the unsupervised setting, the goal is to explore the variance in a single block of data $\mathbf{X}$. For that, a matrix factorization is performed without using any *a priori* knowledge (e.g., no information about the class label of data, the number of classes, etc.), so that natural patterns can be elucidated. This approach is ideal to explore *omics* data in an unbiased fashion, especially in an early phase of the investigation, when no information on molecular species most involved in the process are available [20, 5]. Among unsupervised multivariate analysis tools, Principal Component Analysis (PCA) is undoubtedly the most popular one. However, the interpretation of highly dimensional *omics* data with PCA can often be challenging, and sparse PCA variants [27, 47, 6] are gaining relevance due to their capability to simplify interpretation by performing variable selection within the model calibration.

In the supervised setting, the goal is to explore the variance in a block of data $\mathbf{X}$ that allows the prediction of a response block $\mathbf{Y}$. The latter may contain quantitative data, and therefore we are in the regression domain, or categorical data, and then we are in the classification domain (i.e., control *versus* disease samples). As a result, providing that no overfitting is occurring [46,

2

17], supervised methods can point to the variables (i.e., metabolites, genes,...) that lead to the desired classification. Popular supervised multivariate analysis tools are Partial Least Squares (PLS) regression [19, 46] and its extension to classification PLS Discriminant Analysis (PLS-DA) [2]. Sparse variants of those techniques also exist [30, 7].

All of the above mentioned algorithms are multi-component models that are used together with visualization tools to interpret the data. A popular visualization tool is the scatter-plot of pairs of components, that is, the scores plot and the loadings plot. Interpretation of multivariate plots in *omics* data is limited by the huge dimensionality of data sets, which often requires high numbers of components for proper modeling. An added challenge is the proper selection of model variants [12], which includes the numerous choices of data pre-processing at hand.

In light of these considerations, it is of utmost importance to propose strategies to simplify *omics* data interpretation through multivariate analysis, especially nowadays that multivariate analysis is present in the majority of commercial software packages thus not only for expert chemometricians. To this end, we designed a subspace discriminant index (SDI) aiming to facilitate the inspection of a multi-component model for data exploration. After a supervised or unsupervised multi-component model is built, the SDI can automatically detect the subspace that better discriminates between two or more classes of interest in the observations. Then, with this information, we can visualize scores and loadings corresponding to the selected components. Noteworthy, the SDI also provides of an indication of which model is the most suited to analyze the data set at hand.

The paper is organized as follows. In the next two sections the SDI will be formally introduced and the materials and methods used in the paper discussed. The following three sections demonstrate the use of the SDI in three real world *omics* case studies: toxicology investigations by means of targeted metabolomics, drug safety evaluation by means of untargeted lipidomics, and food science by means of semi-targeted lipidomics. In the last section, we draw the conclusions of the work.

2. The Subspace Discriminant Index

When inspecting an *omics* classification data set of high dimensionality, one is often interested in the characteristics of a specific class or set of classes of observations. However, depending on the data and the model we are using, the information of interest may be hidden in high order components, so that inspection may grow complicated and we may miss relevant detail.

Let us motivate this problem with the example in Fig. 1, which corresponds to the first 2 latent variables (LVs) in a PLS-DA model of a multi-class lipidomics data set we will describe and use later. Even when using a supervised model, which is by nature biased towards the regression/discrimination problem of interest, the first 2 LVs may not be the only subspace to inspect, or even the optimum one. As a matter of fact, in the example, such subspace is
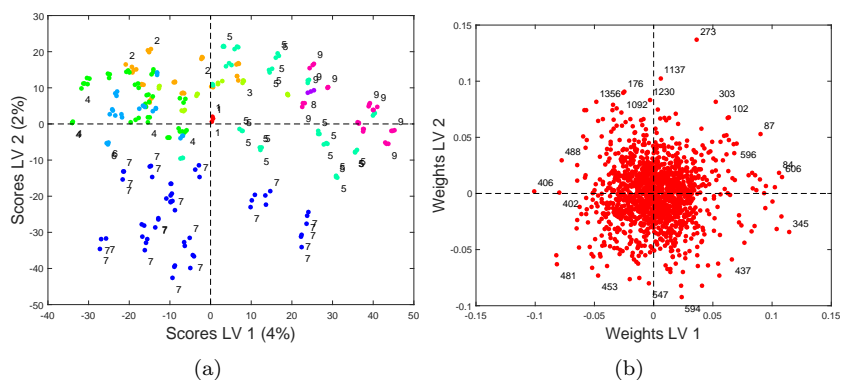
3

Figure 1: Subspace of the first 2 latent variables in the PLS-DA model of the multi-class lipidomics example: (a) scores plot (b) loadings plot.

mainly useful to inspect the characteristics of one single class (number 7) in a 9-classes data set. These characteristics include potential biomarkers associated to the class. Thus, variables 453 (in the same direction in the loading plot to the location of class 7 in the score plot) or 273 (in the opposite direction) may be identified as potential biomarkers (by up-regulation and down-regulation in class 7, respectively). However, one may also desire to see particularities of other classes, and for that the analyst needs to dive into high order components. To give an example, in order to elucidate the characteristics of class 1, with only three observations, one would need to go as far as to inspect the subspace of LV 14 vs LV 15, as shown in Fig. 2, and find out that a biomarker of interest may be variable 1235. Notice there is not a huge difference in variance between subspaces LV 1 vs LV 2 and LV 14 vs LV 15. Still, in the everyday practice of *omics* data analysis, it is unlikely that we get that far by manually inspecting all of the possible pairs of LVs. Note that, in an ordered fashion, this would require to visualize 105 score plots.

One alternative approach to inspecting high order components is to build a number of models in which each class is compared to the rest. This adds the complexity of handling a large number of models to interpret the same data, which can be prone to error and it is specially challenging during initial explorations of the data, where pre-processing and modeling techniques are selected.

To cope with this problem, the SDI can be used to identify the subspace, in a given multi-component projection model, that best discriminates a class (or set of classes) of observations. Then, with this information, we can visualize scores and loadings corresponding to the selected components. All in all, the SDI can be seen as a new addition to the multivariate tool-set useful to facilitate the inspection of a projection model with the ultimate goal of data exploration.

Let us use a general expression for the approximation of the X-block in a multi-component model:
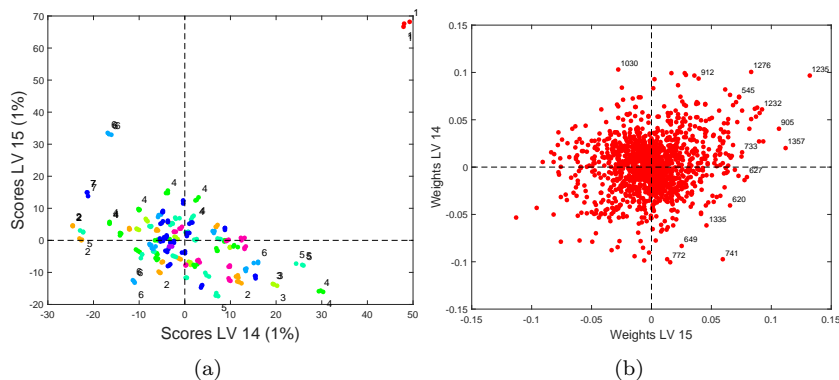
4

Figure 2: Loadings plot corresponding to the subspace of the first 2 latent variables (a) and to the optimum subspace to inspect the characteristics of class1 (b) in the PLS-DA model of the multi-class lipidomics example.

$$\mathbf{X} = \mathbf{U}\mathbf{V}^t + \mathbf{E} \tag{1}$$

where $\mathbf{U}$ are the scores, $\mathbf{V}$ the loadings and $\mathbf{E}$ the residuals. If the number of components A is large, visually inspecting such a model can be a challenging, as already discussed. Alternatively, we define the SDI for a given projection subspace $\mathbb{S}$ and class of observations c, noted as $F_{\mathbb{S}}^c$. The higher the index, the more suitable the subspace $\mathbb{S}$ to inspect the characteristics of class c. Note that in general we are only interested in one-dimensional and two-dimensional subspaces, since those are the potential dimensions of the subspaces we can visualize and rationalize with scores/loadings plots.

The definition of the SDI is grounded on traditional ideas beneath discriminant analysis, t-tests and ANOVA-like models, where the within-group variance is compared to the between-group variance in a ratio:

$$F_{\mathbb{S}}^c = \frac{W_{\mathbb{S}}^c}{B_{\mathbb{S}}^c} \tag{2}$$

where:

$$W_{\mathbb{S}}^c = s_c^2(N_c - 1) - s_{-c}^2(N_{-c} - 1) \tag{3}$$

$$B_{\mathbb{S}}^c = (m_c - m_{-c})^2 \tag{4}$$

and N_c and N_{-c} are the number of observations in the c-th class and outside the c-th class, respectively, s_c and s_{-c} are the corresponding standard deviations and m_c and m_{-c} the corresponding averages.

For a given subspace $\mathbb{S} = a \in \mathrm{R}^1$, so that it corresponds to single component, $W_{\mathbb{S}}^c$, $B_{\mathbb{S}}^c$ and $F_{\mathbb{S}}^c$ are directly computed over the scores of that component: $\mathbf{u}_a$ in the a-th column of $\mathbf{U}$ in eq. (1). To extend this idea to higher order subspaces
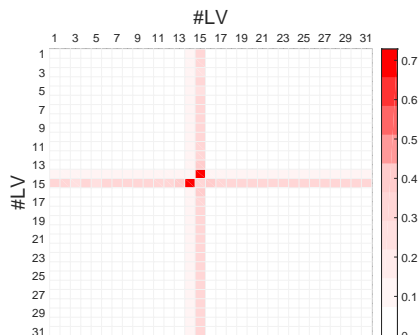
Figure 3: SDI heat map for class 1 in the PLS-DA model of the multi-class lipidomics example.

(e.g. two components), with $\mathbb{S} \in \mathrm{R}^n$ for n above 1, we apply the following strategy. First, we project the scores $\mathbf{U}_{\mathbb{S}}$ onto the most discriminant direction for c in $\mathbb{S}$. Then, we compute $W_{\mathbb{S}}^c$, $B_{\mathbb{S}}^c$ and $F_{\mathbb{S}}^c$ from the projected values. To identify the most discriminant direction in the sub-space, the approach of this paper is to fit a PLS-DA of 1 LV for the specific class of interest against the rest[1]. Thus, the X-block of this PLS-DA are the scores corresponding to subspace $\mathbf{U}_{\mathbb{S}}$, and the Y-block is represented by a dummy variable with 1s for the observations in class c and 0s for the rest. A PLS model with 1 LV is identified from the X-block and Y-block, yielding the regression coefficients $\hat{\mathbf{b}}$ [2]. Subsequently, $W_{\mathbb{S}}^c$, $B_{\mathbb{S}}^c$ and $F_{\mathbb{S}}^c$ are applied over the vector $\mathbf{t}_{\mathbb{S}} = \mathbf{U}_{\mathbb{S}}\hat{\mathbf{b}}$.

One suitable visualization for the SDIs corresponding to a specific model and class of observations is in the form of heat map. Let us come back to the lipidomics example. The SDI heat map for class 1 in the PLS-DA model is shown in Fig. 3. The indices for individual components are located at the diagonal of the heat map, and the indices for 2-component subspaces, with components a and b, are located in the corresponding row-column pair a and b and symmetrically in b and a. The single heat map in the figure assesses the discriminating power of a total of 31 components and $(31 \times 30)/2 + 31 = 496$ different subspaces in the PLS-DA model of the data. From all these possibilities, the figure clearly shows that the maximum SDI for class 1 is found at the subspace of LV 14 vs LV 15, as we anticipated in Fig. 2. Using this plot, the analyst knows where to look.

One comment is in due with respect to multi-class data sets. We may be interested in finding the subspace that best discriminates all the classes on a general basis, instead than a given one. That goal can be achieved by simply combining the SDI maps for the classes, and this can be done in different ways, e.g. with the sum of maps, the normalized sum or the minimum/maximum values. An example is shown in Fig. 4(a), where the normalized sum is used. Since

---

[1]Do not confuse this model with the model over which the SDI is actually computed to enhance interpretation.
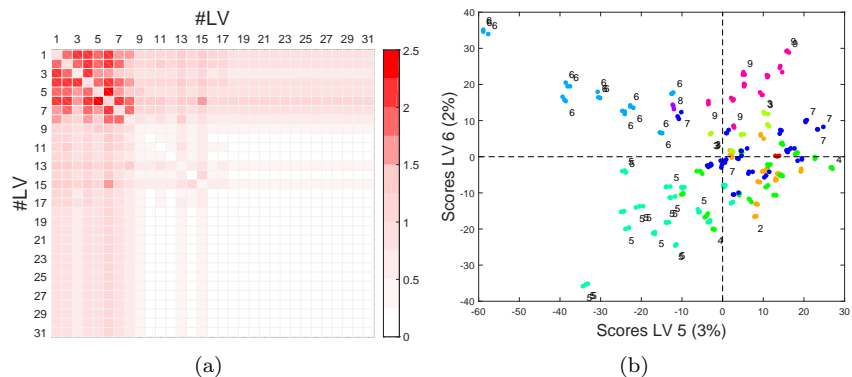
Figure 4: SDI heat map (a) for the normalized sum of classes and the PLS-DA model in Fig. 1(b) and optimum scores plot (b).

the data set is very complex, with several classes, the map has many alternative subspaces of similar SDI. The score plot of the subspace corresponding to the maximum SDI is represented in Fig. 4(b). We can see that the subspace is interesting to study three of the classes, providing more information than the first 2 LVs.

A noteworthy application of an SDI map is to compare different modeling methods in order to select one of them for data inspection. Fig. 5 illustrates the SDI map for PCA and class 1, for comparison with the PLS-DA SDI in Fig. 3. Please, note that we do not normalize the color-scale in the figures. Rather, we compare several maps by looking at the maximum value in the color-scale bar located in the upper right corner. Clearly, PLS-DA, with a maximum SDI above 0.7, is preferable to PCA, with maximum SDI below 0.03. Unlike in PLS-DA, the PCA model splits the discrimination across the PCs, so that the PCA matrix factorization in components is nearly useless to visualize this discrimination. Please, note the discrimination is in the data, not a quality of the model, and we only seek a model to properly visualize and understand it. The SDI shows that PLS-DA is a much more powerful tool to visualize this data than PCA. While this is generally expected, being PLS-DA a supervised approach, it does not necessary need to be the case in all multi-class data sets. In situations where PCA and PLS-DA provide a similar capability for exploration, the former may be preferred due to its unsupervised nature, less prone to over-fitting.

3. Materials & Methods

In the following sections, the results of the application of the proposed SDI using different models are presented and discussed bearing in mind the final objective of simplifying data interpretation. Models considered are PCA, Sparse PCA (SPCA) [47], Group-wise PCA (GPCA) [6] and PLS-DA. When unsupervised approaches show discriminant power, they are generally preferred. For
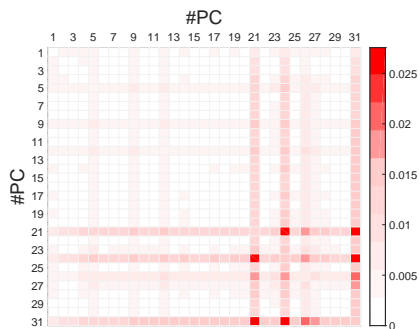
Figure 5: SDI heat map for class 1 in the PLS-DA model of the multi-class lipidomics example.

a fair comparison of different models, given that sparse and supervised models generally capture less variance than classical unsupervised models, the number of components in SDI heat maps span the same amount of variance. We follow [8] to compute scores and variance in sparse models. To select the meta-parameters in SPCA (number of non-zero elements per LV), we use the ckf cross-validation [37]. In GPCA, we select the grouping threshold visually, as suggested in [6].

Three very different real-world metabolomic case studies were selected in order to demonstrate the versatility of the proposed index. The first case study is a two-class problem in which metabolomic analysis is used to understand the effect of a mycotoxyn on wheat. The second case study focuses on the use of lipidomic profiles of drug-treated cells for preclinical drug safety assessment. The third one is a multi-class case study in which lipidomes of olive oils of different nations are compared in order to establish a lipidomic fingerprint that can be used for future geographical localization and contribute to the authenticity control of olive oil products. Details on sample preparations and analysis are given in Supplementary Materials.

The code for the models and the computation of the SDI is freely available in the Matlab MEDA toolbox [9] at the address github.com/josecamachop/MEDA-Toolbox.

4. Two-Class, Targeted Metabolomics Case Study: Mycotoxyn Effect on Wheat

Fusarium head blight (FHB), caused by the fungal plant pathogen *Fusarium graminearum*, is a devastating disease of wheat *Triticum aestivum* causing premature spikelet bleaching shortly after infection [40]. The fungus produces a mycotoxin known as deoxynivalenol (DON) that can have detrimental effects on the health of domestic animals and humans. Therefore, experiments aimed at identifying changes in the metabolome of infected wheat induced by DON were performed with the aim of building models for an early diagnosis of FHB

8

and identifying possible fungicide targets. In this case study, samples of infected wheat were collected as described in Section S1 and the resulted data were explored by means of PCA, SPCA and GPCA. The SDI was subsequently applied to all of the three models, and the heatmaps of the resulting DIs are reported in Fig. 6. Please note that each SDI has the number of components necessary to account for a 50% of the variance in the data. The absolute value of the SDI (0.01 for PCA, above 0.012 for SPCA and GPCA) shows very similar discriminant performance in the models. However, in general sparse variants will be simpler to interpret. PCA shows its best discriminant performance in the subspace LV1-LV2. SPCA and GPCA are optimum in LV1, but GPCA also shows that LV22 may also be inspected.

The corresponding score and loading plots for optimal subspaces are shown in Fig. 7. In PCA, a separation between water mock control and DON-treated samples appears across the first two components, but the loadings are of difficult interpretation since, as commonly happens with *omics* data, many variables have very similar loadings. This limitation is overcome by the SPCA and GPCA models. Indeed, the large amount of zero loadings obtained allows a much more straightforward interpretation: the nonzero loadings were easily identified by Camacho et al. [6] as amino acids, ampholytic amino acids and derivatives suggesting that DON treatment can affect dysregulation of metabolic pathways in which these compounds are involved. The LV22 in GPCA includes an additional variable which scores share a very similar profile to that in LV1, and therefore may also be considered a valid biomarker. Without the SDI, this last biomarker would not have been found.

5. Two-Class, Untargeted Lipidomics Case Study: Drug-Induced Phospholipidosis

Phospholipidosis (PLD) is defined as the abnormal and excessive intracellular accumulation of phospholipids (PLs) within animal cells [22]. In recent years, a number of cationic amphiphilic drugs (CADs) have been reported to cause PLD in humans as side-effect [44, 33, 26]. In light of this, PLD preclinical testing development has become a crucial priority for the pharmaceutical industry. To address this issue a number of combined experimental and in silico tests have been recently proposed [13, 10, 11]. In addition, lipidomics has emerged as powerful tool to directly monitor lipids profile alteration and correlate them with the induced PLD effect [38, 18]. Indeed, lipids that best discriminate between PLD inducer and PLD not inducer drugs could be eventually monitored in preclinical drug safety assessment. Therefore, in the present case study, cells were treated with PLD inducer and not inducer drugs and underwent lipidomic analysis. Samples preparation and data collection are described in Section S1. A total of 33 samples were obtained (Table 1).

The SDI is applied again over PCA, SPCA and GPCA in Fig. 8. The number of components is selected to capture 50% of the variance in the data. The maximum SDI in the GPCA model is one order of magnitude higher that
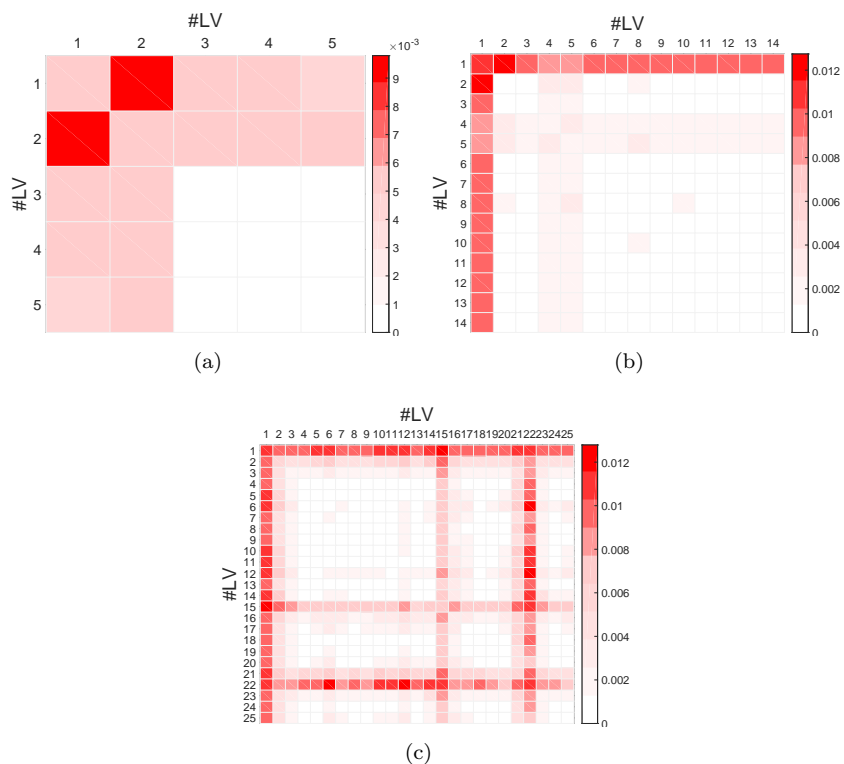
Figure 6: Case Study I: SDI maps of the best subspaces for the (a) PCA, (b) SPCA, and (c) GPCA models. Heat-map scale ranging from white (SDI = 0) to red (maximum SDI within the model, see scale legend at the right side of the maps).

Table 1: Case Study II: Analyzed samples. Drug name and relative PLD induction (inducer: PLD-I; not inducer: PLD-NI), molar concentration (μ M), and number of replicates (N_R).

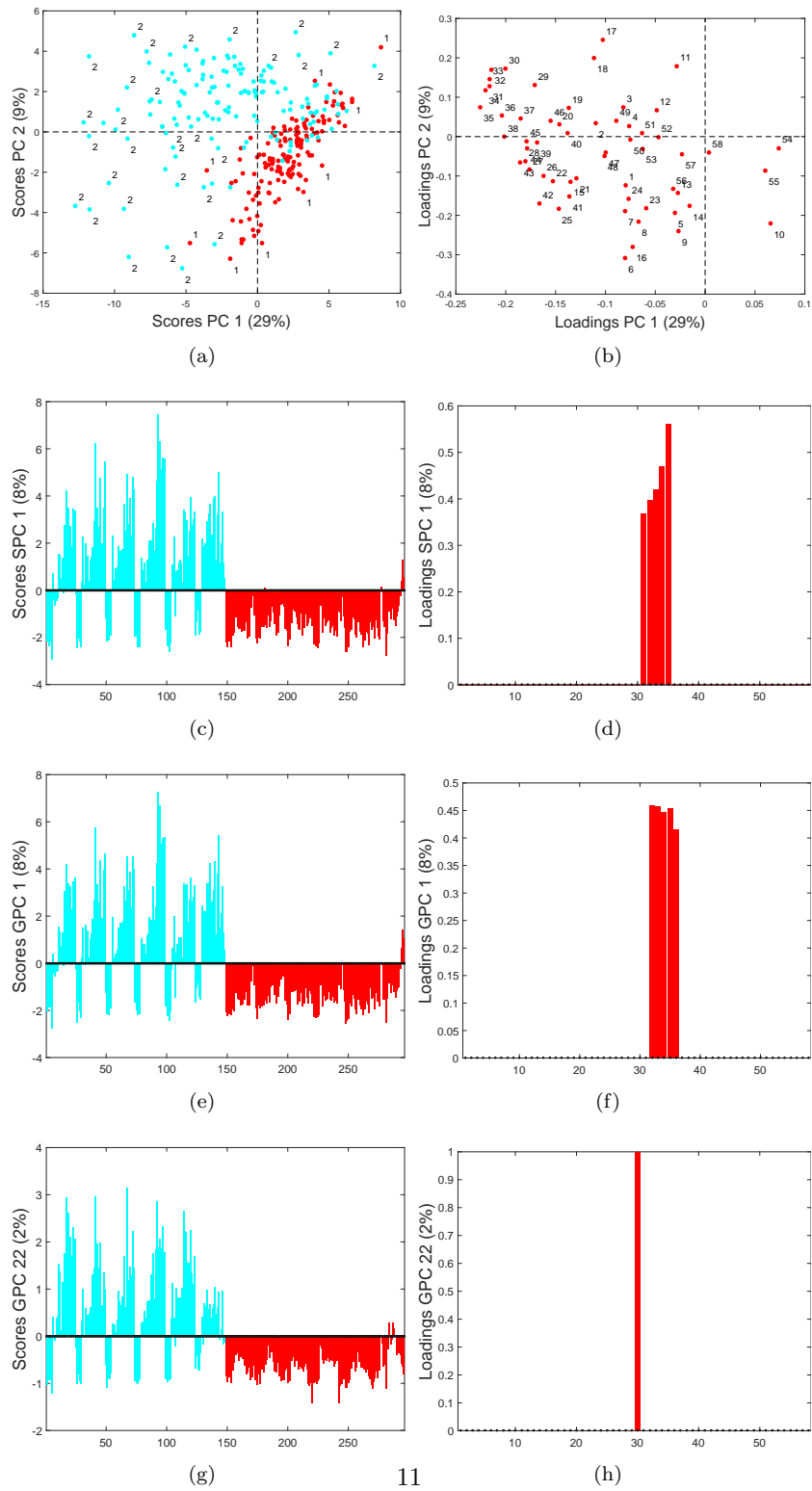| Drug	Effect	Molarity (μ M)	N_R
amiodarone	PLD-I	1	3
		8	3
		12	3
imipramine	PLD-I	1	3
		10	3
		30	3
cimetidine	PLD-NI	4	3
		20	3
		40	3
control	Ctrl	–	2
control-vehicle	Ctrl-V	–	2
blank	Blank	–	2

(a)

(b)

(c)

(d)

(e)

(f)

(g)

11

(h)

Figure 7: Case Study I: score plots (a, c, e, g; cyan - mock water controls, red - DON treated samples) and loading plots (b, d, f, h) of the best subspaces of the PCA, SPCA, GPCA models, respectively.
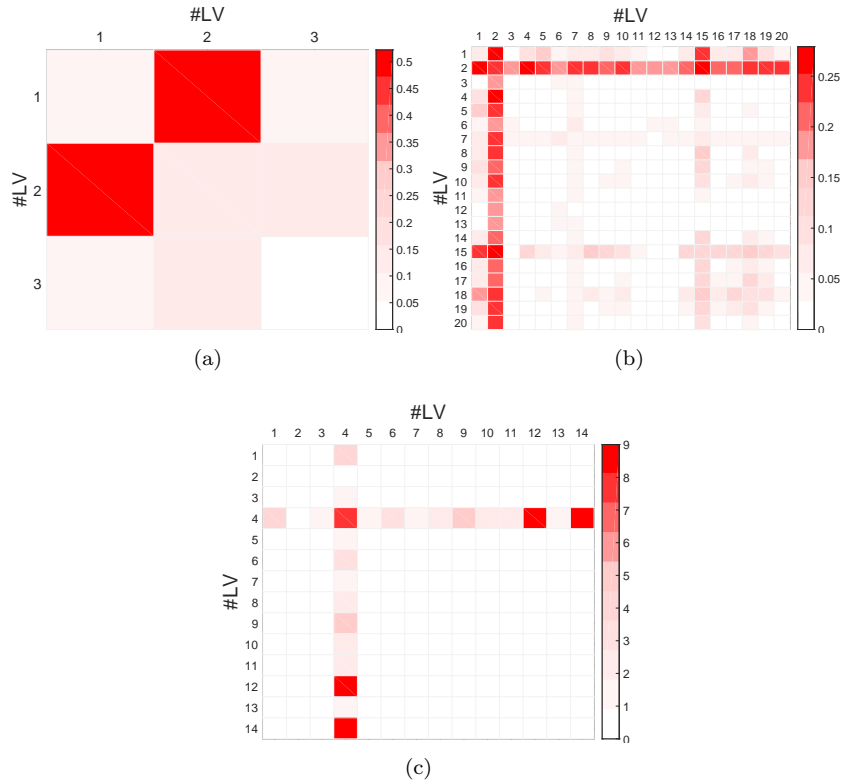
Figure 8: Case Study II: SDI maps of the best subspaces for the (a) PCA, (b) SPCA, and (c) GPCA models. Heat-map scale ranging from white (SDI = 0) to red (maximum SDI within the model, see scale legend at the right side of the maps).

that of SPCA and PCA, respectively. The best subspace in PCA is LV1 vs LV2, in SPCA the second component and in GPCA the 4th one.

The optimum subspaces are compared in Fig. 9. The discrimination in the scores of PCA is very good, but again we have the challenge to explore a crowded loading plot. Biomarkers can be identified approximately around variable 10 (in the direction of the red-scores class) and around variables 45 and 160 (in the other direction). SPCA simplifies this interpretation with less than ten non-zero loadings, around variables 10 and 160, in agreement with PCA, but with a significant loss in discrimination power, as predicted by the SDI. GPCA selects the variables around 45, resulting in a sparser model with an optimum discriminant power. In three plots, the SDI was useful to find the best model and component to interpret the data.
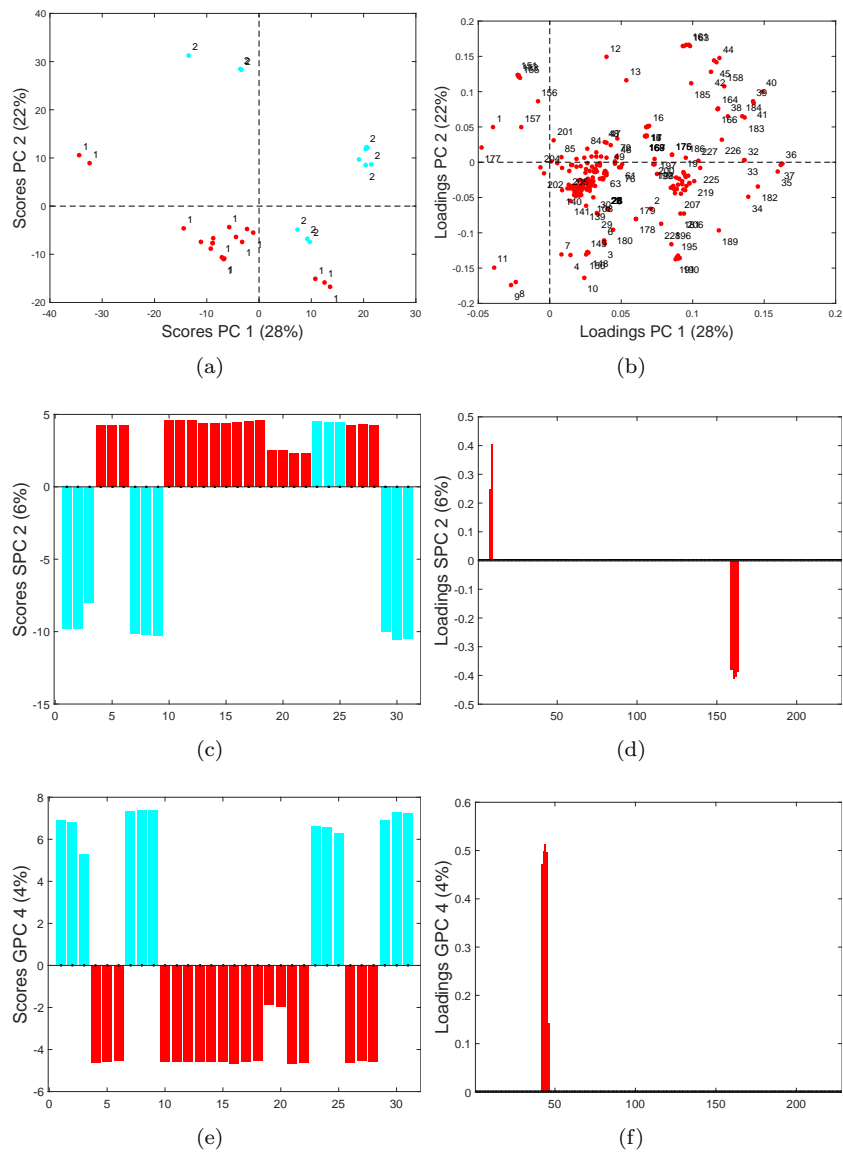
Figure 9: Case Study II: score plots (a, c, e; cyan - PLD-NI, red - PLD-I samples) and loading plots (b, d, f) of the best subspaces of the PCA, SPCA, GPCA models, respectively.

13

Table 2: Case Study III: Analyzed samples. Number (N_S), number of replicates (N_R) and total number of analyzed samples per class (N_{TC}).

Oil Origin	Label	N_S	N_R	N_{TC}
Blanks	-	6	3	18
Cile	1	1	3	3
Croatia	2	11	3	33
Greece	3	6	3	18
Italy	4	15	3	45
Morocco	5	15	3	45
Slovenia	6	8	3	24
Spain	7	17	3	51
Tunisia	8	1	3	3
Turkey	9	9	3	27
			Total	**249**

6. Multi-Class, Semi-Targeted Lipidomics Case Study: Olive Oil Origin

Olive oil is highly valued worldwide for its positive effect on human health. Among the biggest cultivator of olive trees we find Southern Europe, with Italy, Spain and Greece being the most contributing countries, North Africa, and the Near East [25]. From a chemical point of view olive oil, that is the product of mechanical extraction of olive fruit, is a mixture of tri-, di-, and mono-glycerides, as well as phenolic, responsible of its antioxidants activity, and volatile compounds that accounts for its taste and aroma, respectively [14, 4]. The exact mixture composition of olive oil depends on fruit location, variety and ripeness [4]. Therefore, identification and quantitation of these classes of lipids can be used to address several aspects, such as determination of the origin and metabolism of the olive plant, offering a modern tool for quality assessment and authenticity [14, 41]. General guidelines and rules about limits for the composition and the physicochemical parameters of olive oil have been established at the European level [25, 16] in an attempt to protect olive oil quality and authenticity against sophistications and other illegal actions. In this contest, lipidomic fingerprint can be a powerful tool for effectively and unambiguously characterizing olive oils up to the molecular composition. In this case study, 83 olive oils produced from trees cultivated in different countries (Cile, Croatia, Greece, Italy, Morocco, Slovenia, Spain, Tunisia, Turkey, Table 2) were collected and analyzed by means of LC-MS lipidomics with the aim of identifying differences/analogies in their lipidomic fingerprints that could be used for future geographical localization and will contribute to the authenticity control of olive oil products. Each sample was further divided in three aliquots analyzed separately. A total of 249 samples were analyzed as described in Section S1.

For this case study, unsupervised methods did not perform adequately and we had to apply their supervised counterparts. Besides, cross-validation (not
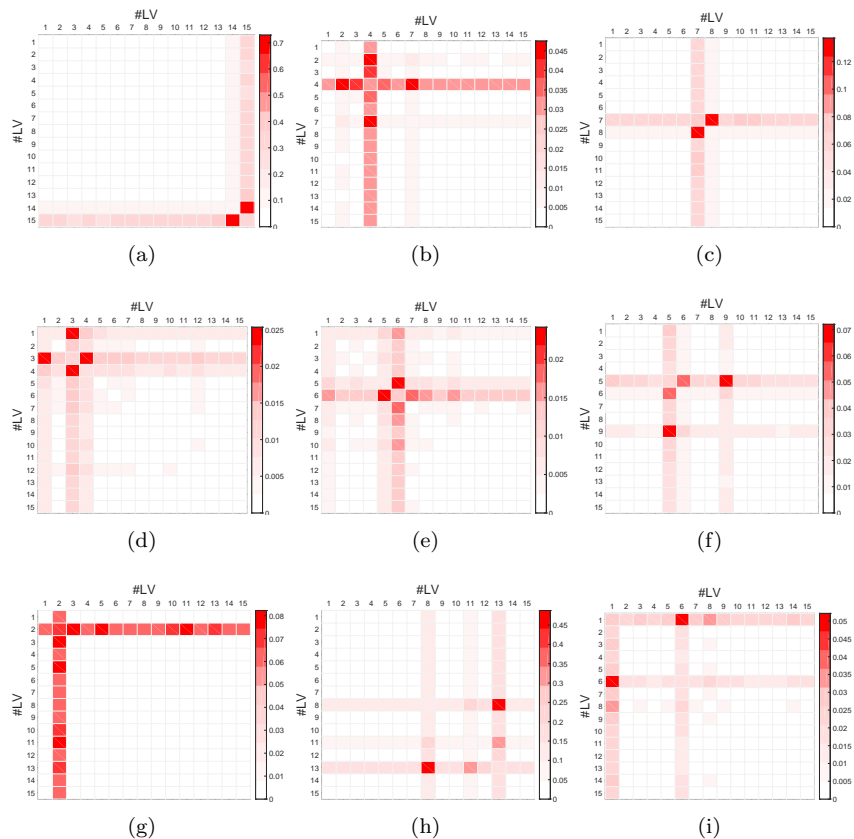
Figure 10: Case Study III: SDI maps for the 1 to 9 classes (a-i, respectively). Heat-map scale ranging from white (SDI = 0) to red (maximum SDI within the model, see scale legend at the right side of the maps)

shown) suggested that sparsity constraints were counterproductive. Therefore, we will focus on the PLS-DA model alone. SDI maps are shown in Fig. 10. We show the number of components that captures 30% of the variance in $\mathbf{X}$ and 95% in $\mathbf{Y}$. In comparison, PCA SDI maps (not shown) present on average one order of magnitude less discrimination performance. Fig. 10 shows that the optimal subspaces are different for different classes, and none of them includes the subspace of the first two LVs. The score plots of optimal subspaces are displayed in Fig. 11, which illustrates the effectiveness of the approach.

7. Conclusions

The synergistic combination of recent advances on the analytical technologies on one side and of the proper data analysis tools on the other side have
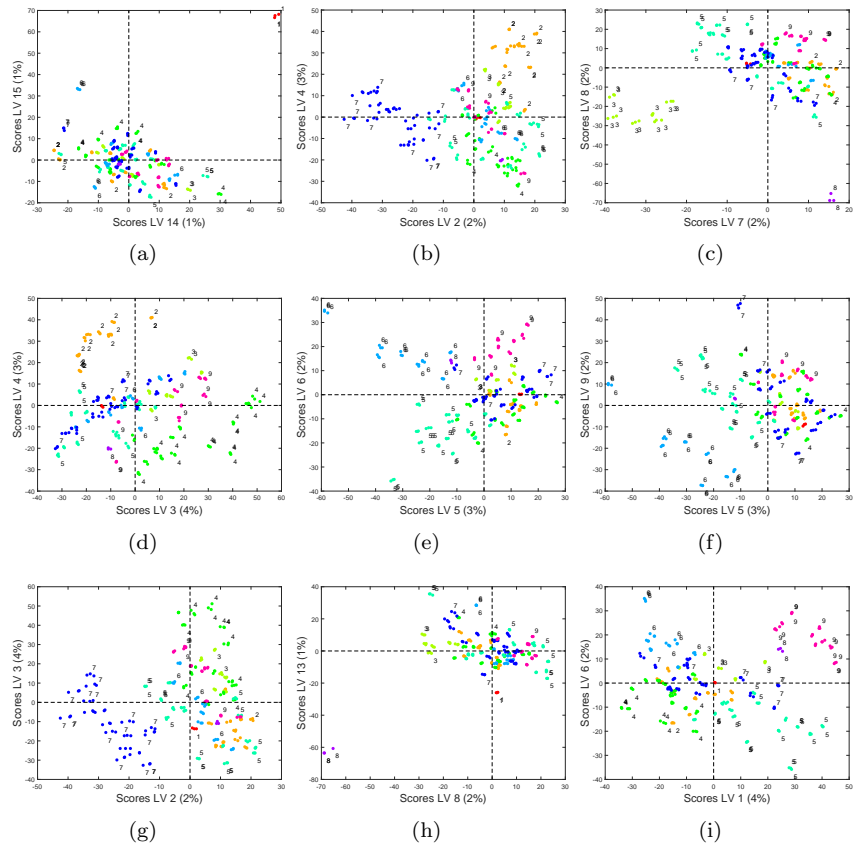
Figure 11: Case Study III: scores plots of optimal subspaces for the 1 to 9 classes (a-i, respectively).

16

boosted the evolution of *omics* sciences. Chemometrics has played and is playing an important role on this, but now that data have increased in both size and dimensionality, and an increasing number of techniques, methods and algorithms have been developed, new tools are needed to simplify data exploration. To address this issue, in the present work we introduced a new subspace discriminant index (SDI) that facilitates the interpretation of a projection model by indicating the best component or pair of components to be inspected for effectively discriminating any class of interest. The efficacy as well as the value of the proposed index was proved in three real world case studies: one in toxicology (two-class), one in pharmaceutical (two-class) and one in food science (multi-class) research areas. In the two-class case studies, we proved that the SDI is an effective tool to guide the exploratory analysis of data with discrimination purposes using unsupervised models. In the multi-class case study, we demonstrated that the SDI is especially useful when the number of classes is relatively high. Noteworthy, the SDI was proven to give an indication on the best choice for the model to explore the data. We believe the SDI will be used by those seeking for an effective still relatively simple way to interpret their *omics* data with multivariate analysis.

Acknowledgement

Compliance with Ethical Requirements

We declare that all authors comply with Springer's ethical policies. No human participants or animals were involved in the study.

S.T. conceived the presented idea. J.C. developed the theoretical formalism, performed the computations, and supervised the findings of this work. M.S. and T.G.T. conceived and planned the experiments for the lipidomic multi-class case study. G.C. conceived and planned the experiments for the two-class case study and contributed to the analysis of the results. All authors discussed the results and contributed to the final manuscript.

[1] Abdullah, L., Evans, J. E., Montague, H., Reed, J. M., Moser, A., Crynen, G., Gonzalez, A., Zakirova, Z., Ross, I., Mullan, C., Mullan, M., Ait-Ghezala, G. and Crawford, F. [2013], 'Chronic elevation of phosphocholine containing lipids in mice exposed to Gulf War agents pyridostigmine bromide and permethrin', *Neurotoxicology and Teratology* **40**, 74–84.

URL: *http://www.sciencedirect.com/science/article/pii/S0892036213001979*
http://linkinghub.elsevier.com/retrieve/pii/S0892036213001979

[2] Barker, M. and Rayens, W. [2003], 'Partial least squares for discrimination', *Journal of Chemometrics* **17**(3), 166–173.
URL: *http://doi.wiley.com/10.1002/cem.785*

[3] Berger, B., Peng, J. and Singh, M. [2013], 'Computational solutions for omics data', *Nature Reviews Genetics* **14**(5), 333–346.
URL: *http://www.nature.com/articles/nrg3433*

[4] Boskou, D., Blekas, G. and Tsimidou, M. [2006], Olive Oil Composition, *in* D. Boskou, ed., 'Olive Oil (Second Edition)', second edi edn, AOCS Press, pp. 41–72.
URL: *https://www.sciencedirect.com/science/article/pii/B9781893997882500080*

[5] Cajka, T. and Fiehn, O. [2016], 'Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics', *Analytical Chemistry* **88**(1), 524–545.
URL: *http://pubs.acs.org/doi/abs/10.1021/acs.analchem.5b04491*

[6] Camacho, J., Rodríguez-Gómez, R. A. and Saccenti, E. [2017], 'Group-wise Principal Component Analysis for Exploratory Data Analysis', *Journal of Computational and Graphical Statistics* .

[7] Camacho, J. and Saccenti, E. [2018], 'Group-wise Partial Least Squares Regression', *Journal of Chemometrics (Wiley)* **32**(3), 1:11.

[8] Camacho, J., Smilde, A., Saccenti, E. and Westerhuis, J. [2020], 'All sparse pca models are wrong, but some are useful. part i: Computation of scores, residuals and explained variance', *Chemometrics and Intelligent Laboratory Systems* **196**, 103907.
URL: *http://www.sciencedirect.com/science/article/pii/S0169743919303636*

[9] Camacho, J., Villegas, A. P., Rodríguez-Gómez, R. A. and Jiménez-Mañas, E. [2015], 'Multivariate Exploratory Data Analysis (MEDA) Toolbox for Matlab', *Chemometrics and Intelligent Laboratory Systems* **143**, 49 – 57.
URL: *http://www.sciencedirect.com/science/article/pii/S0169743915000465*

[10] Ceccarelli, M., Germani, R., Massari, S., Petit, C., Nurisso, A., Wolfender, J.-L. and Goracci, L. [2015], 'Phospholipidosis effect of drugs by adsorption into lipid monolayers', *Colloids and Surfaces B: Biointerfaces* **136**(Supplement C), 175–184.
URL: *http://www.sciencedirect.com/science/article/pii/S0927776515301727*

[11] Ceccarelli, M., Wagner, B., Alvarez-Sánchez, R., Cruciani, G. and Goracci, L. [2017], 'Use of the Distribution Coefficient in Brain Polar Lipids for the Assessment of Drug-Induced Phospholipidosis Risk', *Chemical Research in Toxicology* **30**(5), 1145–1156.
URL: *http://dx.doi.org/10.1021/acs.chemrestox.6b00459*

[12] Checa, A., Bedia, C. and Jaumot, J. [2015], 'Lipidomic data analysis: Tutorial, practical guidelines and applications', *Analytica Chimica Acta* **885**, 1–16.
URL: *http://dx.doi.org/10.1016/j.aca.2015.02.068*

[13] Choi, S. S., Kim, J. S., Valerio, L. G. and Sadrieh, N. [2013], 'In silico modeling to predict drug-induced phospholipidosis', *Toxicology and Applied Pharmacology* **269**(2), 195 – 204.
URL: *http://www.sciencedirect.com/science/article/pii/S0041008X13001063*

[14] Dais, P. and Hatzakis, E. [2013], 'Quality assessment and authentication of virgin olive oil by NMR spectroscopy: A critical review', *Analytica Chimica Acta* **765**, 1–27.
URL: *http://www.sciencedirect.com/science/article/pii/S0003267012017667*
http://linkinghub.elsevier.com/retrieve/pii/S0003267012017667

[15] Ekroos, K. and Ed, R. [2012], *Lipidomics*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.
URL: *http://doi.wiley.com/10.1002/9783527655946*

[16] *European Communities, Regulation 2568/91, Off. J. Eur. Communities 1991, L 248; European Communities, Regulation 1989/2003, Off. J. Eur. Communities 2003, L 295.* [n.d.].

[17] Faber, N. M. and Rajkó, R. [2007], 'How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative', *Analytica Chimica Acta* **595**(1), 98–106.
URL: *http://www.sciencedirect.com/science/article/pii/S0003267007009129*

[18] García-Cañaveras, J. C., Peris-Díaz, M. D., Alcoriza-Balaguer, M. I., Cerdán-Calero, M., Donato, M. T. and Lahoz, A. [2017], 'A lipidomic cell-based assay for studying drug-induced phospholipidosis and steatosis', *ELECTROPHORESIS* **38**(18), 2331–2340.
URL: *http://dx.doi.org/10.1002/elps.201700079*

[19] Geladi, P. and Kowalski, B. [1986], 'Partial least-squares regression: a tutorial.', *Analytica Chimica Acta* **185**, 1–17.

[20] Goracci, L., Tortorella, S., Tiberi, P., Pellegrino, R. M., Di Veroli, A., Valeri, A. and Cruciani, G. [2017], 'Lipostar, a Comprehensive Platform-Neutral Cheminformatics Tool for Lipidomics', *Analytical Chemistry* **89**(11), 6257–6264.
URL: *http://pubs.acs.org/doi/abs/10.1021/acs.analchem.7b01259*

[21] Griffiths, W. J. and Wang, Y. [2009], 'Mass spectrometry: from proteomics to metabolomics and lipidomics', *Chem. Soc. Rev.* **38**(7), 1882–1896.
URL: *http://dx.doi.org/10.1039/B618553N*

[22] H. Halliwell, W. [1997], 'Cationic Amphiphilic Drug-Induced Phospholipidosis', *Toxicologic pathology* **25**, 53–60.

[23] Han, X. [2016], *Lipidomics Comprehensive Mass Spectrometry of Lipids*, Cambridge University Press.
URL: *http://ebooks.cambridge.org/ref/id/CBO9781107415324A009 http://arxiv.org/abs/1011.1669 http://dx.doi.org/10.1088/1751-8113/44/8/085201 http://stacks.iop.org/1751-8121/44/i=8/a=085201?key=crossref.abc74c979a75846b3de48a5587bf708f*

[24] Harkewicz, R. and Dennis, E. A. [2011], 'Applications of Mass Spectrometry to Lipids and Membranes', *Annual review of biochemistry* **80**, 301–325.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410560/*

[25] *International Olive Oil Council, COI/T.15/NC n. 2/rev. 4 (June 6), Madrid, 1996; COI/T.20/Doc no. 24, Madrid, 2001* [n.d.].

[26] J Gonzalez-Rothi, R., Zander, D. and Ros, P. [1995], 'Fluoxetine hydrochloride (Prozac)-induced pulmonary disease', *Chest* **107**, 1763–1765.

[27] Jolliffe, I., Trendafilov, N. and Uddin, M. [2003], 'A modified principal component technique based on the LASSO', *Journal of Computational and Graphical Statistics* .
URL: *http://oro.open.ac.uk/3949/*

[28] Kang, Y. P., Lee, W. J., Hong, J. Y., Lee, S. B., Park, J. H., Kim, D., Park, S., Park, C.-S., Park, S.-W. and Kwon, S. W. [2014], 'Novel Approach for Analysis of Bronchoalveolar Lavage Fluid (BALF) Using HPLC-QTOF-MS-Based Lipidomics: Lipid Levels in Asthmatics and Corticosteroid-Treated Asthmatic Patients', *Journal of Proteome Research* **13**(9), 3919–3929.
URL: *https://doi.org/10.1021/pr5002059 http://pubs.acs.org/doi/10.1021/pr5002059*

[29] Koulman, A., Prentice, P., Wong, M. C. Y., Matthews, L., Bond, N. J., Eiden, M., Griffin, J. L. and Dunger, D. B. [2014], 'The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism', *Metabolomics* **10**(5), 1018–1025.
URL: *http://link.springer.com/10.1007/s11306-014-0628-z*

[30] Lê Cao, K.-A., Rossouw, D., Robert-Granié, C. and Besse, P. [2008], 'A sparse pls for variable selection when integrating omics data', *Statistical applications in genetics and molecular biology* **7**(1).

[31] Li, X., Lu, X., Tian, J., Gao, P., Kong, H. and Xu, G. [2009], 'Application of fuzzy c-means clustering in data analysis of metabolomics', *Analytical Chemistry* **81**(11), 4468–4475. PMID: 19408956.
URL: *http://dx.doi.org/10.1021/ac900353t*

[32] Martin, F.-P. J., Montoliu, I., Collino, S., Scherer, M., Guy, P., Tavazzi, I., Thorimbert, A., Moco, S., Rothney, M. P., Ergun, D. L., Beaumont, M., Ginty, F., Qanadli, S. D., Favre, L., Giusti, V. and Rezzi, S. [2013], 'Topographical Body Fat Distribution Links to Amino Acid and Lipid

Metabolism in Healthy Non-Obese Women', *PLoS ONE* **8**(9), e73445.
URL: *https://doi.org/10.1371/journal.pone.0073445*
http://dx.plos.org/10.1371/journal.pone.0073445

[33] Martin, W. and E Standing, J. [1988], 'Amiodarone pulmonary toxicity: Biochemical evidence for a cellular phospholipidosis in the bronchoalveolar lavage of human subjects', *The Journal of pharmacology and experimental therapeutics* **244**, 774–779.

[34] Nicholson, J. K., Connelly, J., Lindon, J. C. and Holmes, E. [2002], 'INNOVATIONMetabonomics: a platform for studying drug toxicity and gene function', *Nature Reviews Drug Discovery* **1**(2), 153–161.
URL: *http://dx.doi.org/10.1038/nrd728* *http://10.0.4.14/nrd728*
http://www.nature.com/doifinder/10.1038/nrd728

[35] Niemelä, P. S., Castillo, S., Sysi-Aho, M. and Orešič, M. [2009], 'Bioinformatics and computational methods for lipidomics', *Journal of Chromatography B* **877**(26), 2855 – 2862. LIPIDOMICS: DEVELOPMENTS AND APPLICATIONS.
URL: *http://www.sciencedirect.com/science/article/pii/S1570023209000403*

[36] Norris, J. L., Farrow, M. A., Gutierrez, D. B., Palmer, L. D., Muszynski, N., Sherrod, S. D., Pino, J. C., Allen, J. L., Spraggins, J. M., Lubbock, A. L. R., Jordan, A., Burns, W., Poland, J. C., Romer, C., Manier, M. L., Nei, Y.-W., Prentice, B. M., Rose, K. L., Hill, S., Van de Plas, R., Tsui, T., Braman, N. M., Keller, M. R., Rutherford, S. A., Lobdell, N., Lopez, C. F., Lacy, D. B., McLean, J. A., Wikswo, J. P., Skaar, E. P. and Caprioli, R. M. [2017], 'Integrated, High-Throughput, Multiomics Platform Enables Data-Driven Construction of Cellular Responses and Reveals Global Drug Mechanisms of Action', *Journal of Proteome Research* p. acs.jproteome.6b01004.
URL: *http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.6b01004*

[37] Saccenti, E. and Camacho, J. [2015], 'On the use of the observation-wise k-fold operation in PCA cross-validation', *Journal of Chemometrics* **29**(8), 467–478.
URL: *http://dx.doi.org/10.1002/cem.2726*

[38] Saito, K., Ishikawa, M., Yamada, H., Nakatsu, N., Maekawa, K. and Saito, Y. [2016], 'Liver lipid profiling of chemical-induced phospholipidosis', *The FASEB Journal* **30**(1 Supplement), lb498.
URL: *http://www.fasebj.org/content/30/1_Supplement/lb498.abstract*

[39] Schatz, M. C. [2015], 'Biological data sciences in genome research', *Genome Research* **25**(10), 1417–1422.
URL: *http://genome.cshlp.org/lookup/doi/10.1101/gr.191684.115*

[40] Schmale, D. G., Wood-Jones, A. K., Cowger, C., Bergstrom, G. C. and Arellano, C. [n.d.], 'Trichothecene genotypes of gibberella zeae from winter wheat fields in the eastern usa', *Plant Pathology* **60**(5), 909–917.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3059.2011.02443.x*

[41] Shen, Q., Dong, W., Yang, M., Baibado, J. T., Wang, Y., Alqouqa, I. and Cheung, H.-Y. [2013], 'Lipidomic study of olive fruit and oil using TiO2 nanoparticle based matrix solid-phase dispersion and MALDI-TOF/MS', *Food Research International* **54**(2), 2054–2061.
URL: *http://www.sciencedirect.com/science/article/pii/S0963996913005395 http://linkinghub.elsevier.com/retrieve/pii/S0963996913005395*

[42] Szymańska, E., van Dorsten, F. A., Troost, J., Paliukhovich, I., van Velzen, E. J. J., Hendriks, M. M. W. B., Trautwein, E. A., van Duynhoven, J. P. M., Vreeken, R. J. and Smilde, A. K. [2012], 'A lipidomic analysis approach to evaluate the response to cholesterol-lowering food intake', *Metabolomics* **8**(5), 894–906.
URL: *http://link.springer.com/10.1007/s11306-011-0384-2*

[43] Trygg, J., Holmes, E. and Lundstedt, T. [2007], 'Chemometrics in Metabonomics', *Journal of Proteome Research* **6**(2), 469–479.
URL: *http://pubs.acs.org/doi/abs/10.1021/pr060594q*

[44] Tulkens, P. [1986], 'Experimental studies on nephrotoxicity of aminoglycosides at low doses: Mechanisms and perspectives', *The American journal of medicine* **80**, 105–114.

[45] Wang, C., Kong, H., Guan, Y., Yang, J., Gu, J., Yang, S. and Xu, G. [2005], 'Plasma Phospholipid Metabolic Profiling and Biomarkers of Type 2 Diabetes Mellitus Based on High-Performance Liquid Chromatography/Electrospray Mass Spectrometry and Multivariate Statistical Analysis', *Analytical Chemistry* **77**(13), 4108–4116.
URL: *https://doi.org/10.1021/ac0481001 http://pubs.acs.org/doi/abs/10.1021/ac0481001*

[46] Wold, S., Sjöström, M. and Eriksson, L. [2001], 'PLS-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems* **58**(2), 109–130.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0169743901001551*

[47] Zou, H., Hastie, T. and Tibshirani, R. [2006], 'Sparse Principal Component Analysis', *Journal of Computational and Graphical Statistics* **15**(2), 265–286.